CNN-SVM Learning Approach Based Human Activity Recognition

Hend Basly[1](✉), Wael Ouarda[2], Fatma Ezahra Sayadi[3], Bouraoui Ouni[1], and Adel M. Alimi[2]

[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
basly.hend@gmail.com, ouni_bouraoui@yahoo.fr

[2] REGIM-Lab.: REsearch Groups in Intelligent Machines, National Engineering School of Sfax (ENIS), University of Sfax, BP 1173, 3038 Sfax, Tunisia
{wael.ouarda,adel.alimi}@ieee.org

[3] EμE-Lab.: Electronics and Microelectronics Laboratory, Faculty of Sciences of Monastir (FSM), University of Monastir, Environment Avenue, 5019 Monastir, Tunisia
sayadi_fatma@yahoo.fr

Abstract. Although it has been encountered for a long time, the human activity recognition remains a big challenge to tackle. Recently, several deep learning approaches have been proposed to enhance the recognition performance with different areas of application. In this paper, we aim to combine a recent deep learning-based method and a traditional classifier based hand-crafted feature extractors in order to replace the artisanal feature extraction method with a new one. To this end, we used a deep convolutional neural network that offers the possibility of having more powerful extracted features from sequence video frames. The resulting feature vector is then fed as an input to the support vector machine (SVM) classifier to assign each instance to the corresponding label and bythere, recognize the performed activity. The proposed architecture was trained and evaluated on MSR Daily activity 3D dataset. Compared to state of art methods, our proposed technique proves that it has performed better.

Keywords: Convolutional Neural Network (CNN) · Human action recognition · Support Vector Machines (SVM)

1 Introduction

Human Activity Recognition remains a very important research field of numerous computer science organizations because of its potency to provide adapted support for various applications such as human-computer interaction, eHealth

Supported by organization x.

A. El Moataz et al. (Eds.): ICISP 2020, LNCS 12119, pp. 271–281, 2020.
https://doi.org/10.1007/978-3-030-51935-3_29

applications and surveillance. Nowadays, according to the method of feature extraction, the recognition of the human activity system can be classified as a classical or a deep model. A classical model is based on hand-crafted feature descriptors which can be categorized in three types; local features, global features or a combination between them to tackle the human activity recognition problem. The global features designate the image as a whole to describe the entire human body motions. However, the local features are extracted from a set of spatio-temporal interest points (STIPs) to describe the image patches of a human action. Although global methods are able to represent more visual informations by maintaining spatio-temporal structures of the occured actions in the video, they are very sensitive to background variations and partial occlusions. The Local features considers the image as small regions, which is practically computationally expensive. On another side, deep models using deep neural networks are a promising alternative in the image analysis applications areas. Convolutional Neural Network (CNN) is considered as one of the successful deep models for image classification tasks. Traditionally, to deal with such problem of recognition, researcher are obliged to anticipate their algorithms of Human activity recognition by prior data training preprocessing in order to extract a set of features using different types of descriptors such as HOG3D [1], extended SURF [2] and Space Time Interest Points (STIPs) [3] before inputting them to the specific classification algorithm such as HMM, SVM, Random Forest [4–6]. It has been proven that the previous approaches are not very robust due to their poor performance and their requirement in time and memory space. Recently, deep learning architectures are employed in order to change the engineering feature extraction phase by an automatic processing where deep neural networks have been directly applied to the raw data without human intervention to extract deep features. Since the training of a new CNN from scratch requires to load huge amount of data and expensive computational resources, we used the concept of transfer learning and fine tune the parameters of a pretrained model. The initial CNN model was trained on a subset of the ILSVRC-2015 of the large scale ImageNet [7] dataset. Consequently, we decreased the training time, and avoid over fitting by insuring the suitable weight initialization given the quite small used data set. In this study, we proposed an advanced human activity recognition method from video sequence using CNN, where the large scale dataset ImageNet pretrains the network. In fact, a pretrained CNN extracts feature vectors that characterize frames from the raw data. The resulting deep sparse representation of features vectors are fed as input to a multi class support vector machines algorithm to be classified. Since the deep neural networks are more difficult to train, the residual learning approach based ResNet model was proposed to facilitate the training phase. The main contribution of the present work is to propose a learning approach for human activity recognition based CNN and SVM able to classify activities from one shot. The proposed framework is trained and tested on a publicly available dataset, i.e., MSRDailyActivity 3D dataset [8]. Obtained results show that the proposed method outperforms the state-of-the-art methods. The rest of this paper is organized as follows: Sect. 2 highlights some related works, in Sect. 3, we describe our proposed approach. We present the experimental evaluation in Sect. 4. Finally, in Sect. 5, we conclude the paper.

2 Related Works

For Human Activity recognition challenge, an activity has to be represented by a set of features. To represent complex activities, authors in [9] have combined the histogram of oriented gradient (HOG), the motion history image (MHI) and the foreground image (FI). The HOG feature represents the magnitude and the direction of corners and edges, MHI feature is extracted to characterize motion direction and the FI is obtained by background subtraction. Finally, all the resulting features have been merged to be fed as input to a simulated annealing multiple instance learning support vector machine (SMILE-SVM) classifier for human activity recognition. The work of [10] extracted a motion space-time feature descriptor characterizing the video frames by combining the histogram of silhouette and the optical flow values. The first feature is obtained by background subtraction and the second is calculated using the algorithm of Lucas-Kanade [11] inside a normalized bounding box. A multi class SVM classifier has been used to classify the activities. This system was set up to face the restraints of long training time and high dimension of the feature vector. [12] investigates a two distinct stream convNets architecture that includes spatial and temporal networks. In the spatial stream, the action recognition is performed from RGB video frames, whereas in the temporal stream, the recognition of action was made from motion information obtained by stacking dense optical flow between consecutive frames. Both streams are employed as ConvNets and are finally combined by late fusion. Two fusion methods have been considered; a fusion by averaging and a fusion by multi-class linear SVM on softmax scores. The purpose in [13] is to classify the human actions from videos into different classes. The process is performed by extracting interest points from each video, segmenting images and constructing motion history images. After selecting discriminating features and representing images by visual words, a histogram of visual words is elaborated based on features extracted from the motion history images. Finally, the extracted features vectors are used to train a support vector machine for action classification. [14] proposed a system to recognize abnormal comportment providing an alert to the accurate user on his android mobile phone. The task is to extract features using Scale Invariant Feature Transform (SIFT) descriptor for each video after dividing them into number of frames. The extracted features are then exploited as input to two different types of classifiers, i.e; the K Nearest Neighbor (KNN) and the Support Vector Machine (SVM) to classify the actions.

3 Proposed Approach

3.1 Convolutional Neural Networks (CNN)

As recent written works [12,24,27] has proven, the deep hierarchical visual feature extractors are currently outperforming traditional hand-crafted descriptor, and are more generalizable and accurate when dealing with important levels of immanent noise problems. To describe the activities in a frame-wise way, we chose to use the CNN approach based on RGB data because of its widespread application in different areas. CNNs are also advantageous by their reduction of the number of parameters and connections used on artificial neural model to facilitate their training phase. In this step, the question now is how to represent the human actions in each extracted frame of the video. To extract the most pertinent and significant features from the raw RGB video frame, we employed a pre-trained deep CNN architecture with pre-trained parameters based on ImageNet. The original CNN was trained on the 1.2M high-resolution images of the ILSVRC2015 classification training subset of the ImageNet dataset. Though, in the proposed method, we used a deep CNN network architecture to generate a probability vector for each input frame which represents the probability of the presence of the different objects present in each individual frame. A ResNet model is used with pre-trained parameters from ImageNet database and applied to extract sparse and pertinent residual representations of features from video frames of each sequence video. 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 network takes an input of size 224 × 224 which was reduced five times in the network by a stride of 2. The output obtained from the average pooling operation is applied to the final feature map of the network followed by the fully connected layer. The resulting vector from the last pooling layer is considered as the features representation generated from the reused pretrained model in a feedforward pass. After each convolution, a batch normalization and an ReLU are achieved. The residual units are represented as:

$$x_{l+1} = f(x_l + F(x_l; W_l)) \tag{1}$$

where x_l and x_{l+1} correspond to the input and the output of the $\mathrm{l}^t h$ layer, F denotes a nonlinear residual mapping characterized by convolutional filter weights Wl and f corresponds to the ReLU function. The main advantage of handling residual units in such types of networks, is that their skip connections or "shortcuts" allow the direct propagation of signals over all the network' layers. This design is very advantageous mainly during the backpropagation phase; in fact, gradients are directly propagated from the loss layer to all the other preceding layers while skipping some intermediate layers which have the potential to provoke the deterioration or the disappearance of the gradient signal. This strategy helped the network to appreciate the accuracy gained from deeper architectures. Since training a new deep CNN model from scratch requires important loads of data and elevated resources of computation, we have implemented a transfer learning procedure to fine-tune the parameters of a pre-trained model. We adopted an original CNN model that was pretrained on a subset of the large-scale image classification dataset such as the ImageNet. Proceeding in this way, we succeed to reduce the required time for training and to avoid our dataset from overfitting by assuring a good initialization of weights, given the quiet small available dataset. In fact, the dataset was artificially augmented by using three techniques. First random reflect frames in the left direction, second a random horizontal translation that consists of moving frames along the horizontal direction, and finally, a random vertical translation is applied by moving frames on the vertical direction. In reality, the last layer of the adopted CNN model is a classification layer; though, in the present study, we removed this layer and exploited the output of the preceding layer as frame features for the classification step. Instead of the eliminated layer, the SVM classifier has been employed to predict the human activity label. Figure 1 summarizes the architecture of the proposed action recognition model.

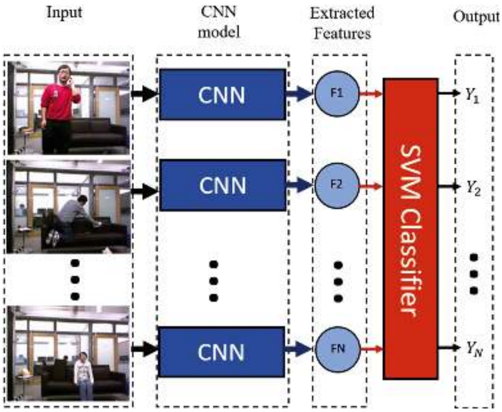


Fig. 1. Architecture of the proposed action recognition model.

3.2 Support Vector Machines (SVM)

SVM is supposed as machine learning classifier method that gives good results in comparison with other types of classifier. We decided to use it in this study because of its effectiveness when dealing with quiet small datasets and its performance in high dimensional spaces [15,16,25–29]. The principal idea behind the use of SVM is to applicate a supervised learning algorithm facilitating to find the optimal hyperplane that separates the feature space. During training, the SVM generates hyperplanes in a high dimensional space to separate the training dataset into different classes. If the training data subset are not linearly separable, a kernel function SVM is used to transmit the data to a new vector space. SVM performs well with large scale training datasets and yields to accurate and effective results. For a given training dataset; $D(x_1, y_1), (x_2, y_2), ...(x_N, y_N)$ where $x_i \in \mathbb{R}^n$ and memberships $y_i \in \pm 1$ classes; i represents the label corresponding to each action in the defined dataset. To determine a decision function for a linear classification, the hyperplane separation is represented by:

$$y_i = sng((w\Delta x_i) + b) \tag{2}$$

A generic hyperplane is defined by satisfying the condition:

$$w \cdot x_i + b = 0 \tag{3}$$

When delimited by margins, the set of hyperplanes can be written as:

$$y_i \cdot ((w \cdot x_i) + b) \geq 1 \tag{4}$$

To formulate the optimal hyperplane that separates the data, we should minimize:

$$1/2\|w\|^2 \tag{5}$$

Subject to the constraints of Eq. 4).

Multi-class SVM. Even though SVM were initially developed for binary classification, it can be successfully extended to be applied to multiclass classification problems. The main strategy consists to separate the multiclass problem into many biclass problems and combine the outputs of all the sub-binary classifiers to provide the final class prediction of a sample. Fundamentally, there are two main methods for multiclass SVM. The first type is called "oneagainstone" [17], it consists to construct one classifier per pair of classes and combine binary classifiers in a way to form a multi-class classifier by selecting the most voted class. So, $N(N-1)/2$ binary SVM classifiers are needed, each of them is trained on the samples of the two corresponding classes. The second method is called "oneagainstall" [27] and it considers all the classes of data in one optimization problem. In fact, for each classifier, the considered class is fitted against all the other classes, so, N number of classes use N SVM classifiers. When using the latter technique, the training process takes a long time.

4 Dataset and Evaluation Procedure

4.1 MSRDailyActivity 3D Dataset

The"MSRDailyActivity 3D"dataset [12] is an RGB sequences dataset that contains sixteen daily human activities. The database was captured by a kinect camera around various objects, and the humans in question are located at different distances from the camera. Activities are accomplished by ten different subjects, the most of them are categorized as "human object interactions". Activities were performed twice by each person in two different positions; i.e; the "standing" and the "sitting" situation.

4.2 Implementation Details

The deep CNN model was trained using Matlab 2018. Our approach based CNN model was performed on a machine equipped with a NVIDIA GeForce 960M GPU, 64 GB memory and an Intel Core i7-6700 HQ (2.60 GHz) processor. Our dataset was artificially augmented. This technique allows to avoid the problem of dataset overfitting. Each video from our dataset were split into frames which serve as input to the pre-trained CNN model. In the training stage, a 224 × 224 frame is randomly reflected from the selected frame; it then undergoes a random horizontal and vertical translation. These operations are applied in such a way that the training dataset is augmented at each iteration. The 2048 dimensional vector resulting from the last pooling layer of the ResNet model were used to activate the training and testing subsets. The resulting vectors were used as training and test data for the multi-class SVM classifier. The training process is performed using a mini-batch stochastic gradient descent with a momentum set to 0.9 to learn the network weights. At each iteration, a mini-batch size of 50 samples is constructed by sampling the training videos by 50, from which a single frame is selected randomly. During our experimentation, the learning rate is initially set to 1e^{-4} and the network is trained for 6 epochs. We also tried to increment the number of epochs but we got always overfitting. For our used multi–class SVM classifier, we chose to employ the linear function kernel to project the original linear or nonlinear dataset into a higher dimensional space in order to make it linearly separable and to give a better performance for the SVM. The Linear kernel is a simple kernel function based on the penalty parameter C described by the following format:

$$K(x, y) = x^T y + c \tag{6}$$

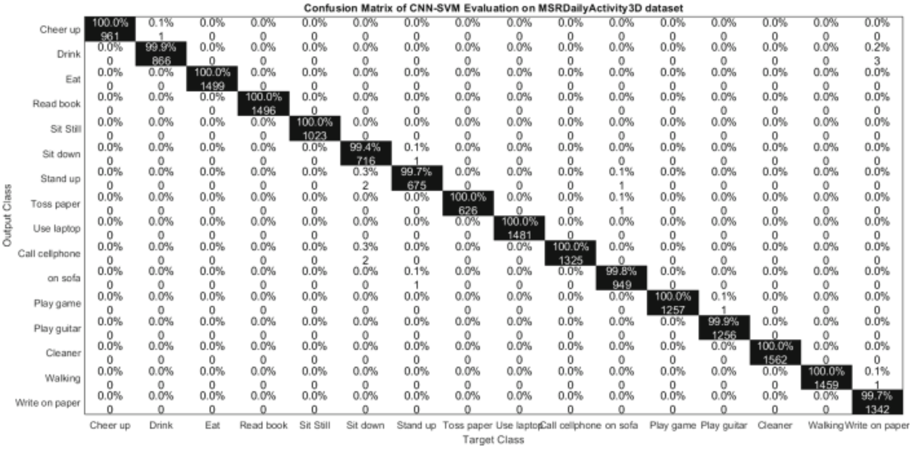
4.3 Evaluation Methodology

During experimentation, we evaluated our method on the dataset described above: 70% used for the training stage and 30% from data are used for testing. Firstly, each frame is resized to 224 × 224 resolution. We have determined the confusion matrix of our proposed system in order to demonstrate the correspondence between the predicted labels along the x-axis and the true labels along the y-axis and to represent the recognition performance for each action class in the MSRDailyActivity 3D dataset. Generally, a confusion matrix involves four groupings: TP (True positive) mean the instances that are correctly identified as positives, FP (False positive) refers to the negative examples incorrectly identified as positive, TN (True negative) refers to the negative instances that are correctly predicted as negative, and FN (false negative) represents the positive instances incorrectly predicted as negative. We also evaluate different performance metrics of our proposed approach by calculating the precision, recall and f-measure values as shown in Table 1.

Table 1. Performance metrics results obtained with our proposed approach.

Accuracy (%)	Precision (%)	Recall	F-measure
99.92	98.77	99.79	99.28

Figure 2 demonstrates that the most confusion is between sit down and stand up labels. This misclassification can be explained by the similarity in a few steps when carrying out both of actions which contain a person in a half-sitting position. İn fact, the middle frames of the two classes sit down and stand up presenting a person in a half setting position are making the confusion, because of their repetition in the two cases. Whereas more than half of the classes have been correctly classified at 100%.



Confusion Matrix of CNN-SVM Evaluation on MSRDailyActivity3D dataset

Output Class

	Cheer up	Drink	Eat	Read book	Sit Still	Sit down	Stand up	Toss paper	Use laptop	Call cellphone	on sofa	Play game	Play guitar	Cleaner	Walking	Write on paper
Cheer up	100.0% 961	0.1% 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	0.0% 0
Drink	0.0% 0	99.9% 866	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	0.2% 3
Eat	0.0% 0	0.0% 0	100.0% 1499	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
Read book	0.0% 0	0.0% 0	0.0% 0	100.0% 1496	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
Sit Still	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 1023	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
Sit down	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	99.4% 716	0.1% 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
Stand up	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.3% 2	99.7% 675	0.0% 0	0.0% 0	0.0% 0	0.1% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Toss paper	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 626	0.0% 0	0.0% 0	0.1% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Use laptop	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% 1481	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Call cellphone	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.3% 2	0.0% 0	0.0% 0	0.0% 0	100.0% 1325	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
on sofa	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.1% 1	0.0% 0	0.0% 0	0.0% 0	99.8% 949	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Play game	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% 1257	0.1% 1	0.0% 0	0.0% 0	0.0% 0
Play guitar	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	99.9% 1256	0.0% 0	0.0% 0	0.0% 0
Cleaner	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% 1562	0.0% 0	0.0% 0
Walking	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	0.0% 0	100.0% 1459	0.1% 1
Write on paper	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	0.0% 0	0.0% 0	99.7% 1342

Target Class

Fig. 2. Confusion matrix of the learning CNN–SVM approach evaluation on MSRDailyActivity3D dataset.

Table 2 notices that our approach has achieved a good recognition performance and outperforms other state-of-the-art methods on MSRDailyActivity 3D dataset. Achieved performance confirms the generalization competence of our learned representations across domains. The work of [18] has obtained bad results in this dataset despite it was based on the combination of two deep neural network models which are the CNN and LSTM. Whereas the implemented CNN model for feature extraction is not based transfer learning concept. Based on all these observations, we can deduce that pretraining a model on a large-scale dataset and fine tune his hyper-parameters on a small one is very efficient to obtain good performance rate. We have also combined the same pretrained ResNet model which was used to extract features, once with a Multi Layer Perception (MLP) classifier and another time with a Long Short Term Memory (LSTM) network. The obtained results show that using a multi-class SVM classifier, gives the best result.

In order to investigate on the effect of the choice of the SVM kernel, we have performed a classification using Radial Basis Function (RBF) kernel. The results were not interesting due to the relevance of the feature representation obtained from convolutional neural network.

Table 2. Comparison of different HAR approaches on MSRDAily Activity 3D dataset.

Reference	Method	Accuracy (%)
Sial et al. [19]	D-STIP + D-DESC+ RGB-DESC +SVM	92.00
Nunez et al. [18]	CNN + LSTM	63.10
Lu Xia et al. [20]	DSTIP + DCSF + SVM	96.70
Wang et al. [21]	LOP + FTP + AEM +SVM	85.75
Shahroudy et al. [22]	Dense trajectories with HOG, HOF and MBH + skeleton joints + SVM	91.25
Wang et al. [23]	WHDMM + Three channel 3D ConvNet	85.00
LAHAR-CNN (Ours)	Pretrained CNN + MLP	99.40
DTR-HAR (Ours)	Pretrained CNN + LSTM	91.56
Our approach	Pretrained CNN + SVM	99.92

5 Conclusion

In this study we presented the support vector machines approach for human activity recognition task. We proposed to use a pre-trained CNN approach based ResNet model in order to extract spatial and temporal features from consecutive video frames. Our proposed architecture was trained and tested on MSRDaily-Activity 3D dataset and it achieved a good recognition performance. For our future works, we propose to use a combination of a genetic algorithm with support vector machines in order to optimize the weights of the used CNN model leading to automatically improve the performance. Likewise, we would like to expend the proposed model for more large-scale dataset such as NTU RGB+D because the used dataset is small and the used pretrained CNN model can be more effective when applied to a big one.

References

1. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: 19th British Machine Vision Conference, BMVC 2008 on Proceedings, pp. 275:1–10. BMVA Press, Leeds (2008)
2. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88688-4_48
3. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Proceedings. 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), pp. 1:65–72. IEEE, Beijing (2005)
4. Zhu, C., Sheng, W.: Multi-sensor fusion for human daily activity recognition in robot-assisted living. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, pp. 303–304. ACM (2009)
5. Ghosh, A., Riccardi, G.: Recognizing human activities from smartphone sensor signals. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 865–868. ACM (2014)

6. Rahman, S.A., Merck, C., Huang, Y., Kleinberg, S.: Unintrusive eating recognition using Google Glass. In: Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth 2015, pp. 108–111. Institute of Electrical and Electronics Engineers Inc, United States (2015). https://doi.org/10.4108/icst.pervasivehealth.2015.259044
7. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
8. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290–1297. IEEE, Providence (2012)
9. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: Proceedings of 12th IEEE International Conference on Computer Vision, pp. 128–135. IEEE, Kyoto (2009)
10. Chathuramali, K.M., Rodrigo, R.: Faster human activity recognition with SVM. In: International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 197–203. IEEE, Colombo (2012)
11. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI 1981), pp. 674–679. Morgan Kaufmann Publishers Inc., Vancouver (1981)
12. 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, Montreal (2014)
13. Siddiqui, S., Khan, M.A., Bashir, K., Sharif, M., Azam, F., Javed, M.Y.: Human action recognition: a construction of codebook by discriminative features selection approach. Int. J. Appl. Pattern Recogn. **5**(3), 206–228 (2018)
14. Kale, G.: Human activity recognition on real time and offline dataset. Int. J. Intell. Syst. Appl. Eng. **7**(1), 60–65 (2019)
15. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
16. Ikram, S.T., Cherukuri, A.K.: Improving accuracy of intrusion detection model using PCA and optimized SVM. J. Comput. Inf. Technol. **24**(2), 133–148 (2016)
17. Xu, Y., Zomer, S., Brereton, R.G.: Support vector machines: a recent method for classification in chemometrics. Critical Rev. Anal. Chem. **36**(3–4), 177–188 (2006)
18. 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 Recogn. **76**, 80–94 (2018)
19. Sial, H.A., Yousaf, M.H., Hussain, F.: Spatio-temporal RGBD cuboids feature for human activity recognition. Nucleus **55**(3), 139–149 (2018)
20. Xia, L., Aggarwal, J.K.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2834–2841. IEEE Computer Society, USA (2013)
21. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3D human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **36**(5), 914–927 (2013)
22. Shahroudy, A., Ng, T.T., Yang, Q., Wang, G.: Multimodal multipart learning for action recognition in depth videos. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 2123–2129 (2015)

23. Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., Ogunbona, P.O.: Action recognition from depth maps using deep convolutional neural networks. IEEE Trans. Hum.-Mach. Syst. **46**(4), 498–509 (2015)
24. Jarraya, I., Ouarda, W., Alimi, A.M.: Deep neural network features for horses identity recognition using multiview horses' face pattern. In: Ninth International Conference on Machine Vision (ICMV), pp. 103410B. International Society for Optics and Photonics (2016)
25. Sassi, A., Ouarda, W., Ben Amar, C., Miguet, S.: Neural approach for context scene image classification based on geometric, texture and color information. In: Chen, L., Ben Amor, B., Ghorbel, F. (eds.) RFMI 2017. CCIS, vol. 842, pp. 110–120. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-19816-9_9
26. Nasri, H., Ouarda, W., Alimi, A.M.: ReLiDSS: novel lie detection system from speech signal. In: IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), pp. 1–8. IEEE (2016)
27. Larbi, K., Ouarda, W., Drira, H., Amor, B.B., Amar, C.B.: DeepColorFASD: face anti spoofing solution using a multi channeled color spaces CNN. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 4011–4016. IEEE, Miyazaki (2018)
28. Ehaimir, M.E., Jarraya, I., Ouarda, W., Alimi, A.M.: Human gait identity recognition system based on gait pal and pal entropy (GPPE) and distances features fusion. In: Sudan Conference on Computer Science and Information Technology (SCCSIT), pp. 1–5. IEEE, Elnihood (2017)
29. Ghabri, S., Ouarda, W., Alimi, A.M.: Towards human behavior recognition based on spatio temporal features and support vector machines. In: Ninth International Conference on Machine Vision (ICMV), pp. 103410E. International Society for Optics and Photonics (2016)