AUT Journal of Electrical Engineering

AUT J. Elec. Eng., 53(1) (2021) 17-26
DOI: 10.22060/eej.2020.18517.5357

A Comparison of Two Neural Network Based Methods for Human Activity Recognition

S. zebhi[1], S. almodarresi[2*], V, abootalebi[3]

1PhD candidate, electrical engineering department, yazd university, yazd, iran
2Associate professor, electrical engineering department, yazd university, yazd, iran
3Associate professor, electrical engineering department, yazd university, yazd, iran

ABSTRACT: In this paper, two different methods of human activity recognition based on video signals are introduced. The first method explores the effectiveness of combining feature descriptors obtained by local descriptors and artificial neural network classifier. It is used in the traditional approach and the local descriptors extract interest points or local patches from the videos, and the feature vectors are later constructed based on the intrests, and eventually feature vectors are used as the input of a two-layer feed-forward artificial neural network (ANN). Experimental results show that using the HOG3D descriptor with ANN gives the best performance. On the other hand, deep learning architectures have attracted much consideration for automatic feature extraction in the last years, so an improved 3D convolutional neural network architecture is also designed as the second method. They are implemented and compared with state-of-the-art approaches on two data sets. The results exhibit that method 1 is superior when the shortage of sample data is the main restriction. It respectively achieves recognition accuracies of 97.8% and 99.8% for the Weizmann and KTH action data sets. In addition, method 2 is considerable for its automatic features extraction, and achieves an acceptable result with lots of original training data. As a result, it gets recognition accuracy of 92% for the KTH data set while this value is drastically reduced for the Weizmann data set.

Review History:

Received: May, 31, 2020
Revised: Sep. 17, 2020
Accepted: Oct. 14, 2020
Available Online: Jun. 01, 2021

Keywords:

Local descriptors
Artificial neural network
3D convolutional neural network
Histogram of oriented gradients 3D (HOG3D)

1- INTRODUCTION

The aim of activity recognition is to identify human activities in real-life scenarios. Correct activity recognition is challenging to achieve since human activity is complicated and very different from each other. There are many studies reported for different usages based on human activity recognition, especially sport activity, pedestrian traffic, healthcare applications and etc. These approaches [1–4] can be split into traditional methods with their hand-crafted features and deep learning architectures with their automatic feature extraction.

Local descriptors are applied to discriminate the human activity videos in the traditional methods. These effective descriptors concentrate on special local patches that are defined by interest point detectors or densely sampling [5] and robust to partial occlusion, rotation and background noise [6].

Laptev [7] showed 3D Harris corner by calculating local, spatiotemporal N-jets as the descriptor. The descriptor is scale-invariant since they estimate the spatiotemporal extents of detected events by maximizing a normalized spatiotemporal laplacian operator over spatial and temporal scales. In addition, the proposed descriptors are demonstrated to be robust to occlusions and dynamic cluttered backgrounds in the human motion analysis.

Lowe proposed the scale-invariant feature transform (SIFT) [8]. It is broadly applied in local presentation for its scale and rotation invariance, as well as the robustness to affine distortion, variations in 3D viewpoint, addition of noise and variation in illumination. Scovanner et al., [9] presented a 3D SIFT descriptor and applied it on human activity recognition. The 2D gradient magnitude and orientation are expanded in 3D formulation; therefore, forming the sub-histograms that code the 3D SIFT descriptor. As a result, videos are defined as a bag of spatiotemporal words using the 3D SIFT descriptor. In addition, a feature grouping histogram, that groups the co-occurred words out of the original, is utilized to make a more distinctive action video representation, and applied for classification.

The speeded-up robust features (SURF) [10] method is a scale and rotation invariant detector and descriptor. The most significant feature of SURF is the advancement of performance comparing to the prior method. In the interest point detection, the method uses a strategy that analyzes the input image at different scales to guaranty invariance to scale variations. Uijlings et al., [11] proposed the histogram of oriented gradients (HOG) and histogram of flow (HOF) descriptors. Klaser et al., [5] generalized the HOG descriptor to videos and introduced the HOG3D. Integral images are expanded to integral videos for effective 3D gradient

*Corresponding author’s email: smta@yazd.ac.ir

computation.

Guha et al., [12] defined an efficient approach for finding and calculating main motion templates. The features are designed to acquire the distinctive, scale-invariant region that has important information about the local changes of the signal along both spatial and temporal dimensions, that are the Local Motion Pattern (LMP) descriptors.

In the traditional approaches based on local descriptors, extracted features were classified using parameters optimization [5], support vector machines (SVM) [11, 13-14], or random sample reconstruction (RSR) [12]. Abdelbaky et al., [3] calculated several short-time motion energy images to acquire motion information. Hierarchical local motion features are learned from these images by the deep design of PCANet, and their dimension is decreased by applying the principal component analysis algorithm. Eventually linear SVM is used for classifying the extracted features. Khan et al., [1] accomplished recognition in two main stages. Preprocessing and segmentation for extracting saliency map are done in the first stage. In addition, classifying is performed by neural network on the extracted features in the second stage. In [2], view-invariant features are acquired by extracting the holistic features from different temporal scales. Later, three operational systems are evaluated for classifying features. Singh et al., [4] applied discrete wavelet transform and SVM classifier for recognition. Scale and color normalization are applied in the first step. Background subtraction is applied in the next step to remove any extra affect in the background. Later, wavelet features are extracted from all the resulted frames of video for the classification step. In section [15], Maclaurin coefficients of the density function and moments of the distribution are used to encode the distribution of the video descriptors.

On the other hand, there has recently been an increasing interest in deep designs methods, for their automatic feature construction process. One of the most common deep designs is the convolutional neural network (CNN) architecture. This model extracts features from both the spatial and the temporal dimensions by doing 3D convolutions, and acquiring the motion information coded in several neighbor frames [16]. Ji et al., [17] produced multiple channels of information from the input frames by performing 3D convolutions. They demonstrated adding optical flow as another input modality can considerably enhance the recognition performance. Therefore, four different channels of information were generated by optical flows and gradients in the horizontal and vertical directions from each frame to apply to 3D CNNs [18]. Similarly, in [19], the 3D CNN was constructed to process multi-channel features that are produced by processing the training samples through the gray scale, improved L-K optical flow and Gabor filter. Liu et al., [20] proposed Temporal Convolutional 3D Network (T-C3D) for real-time processing, that the temporal dynamics of the whole video are explored with a new temporal encoding method.

In this study, two different methods for activity recognition are proposed. Method 1 is categorized into traditional approaches and consists of four steps. Interest points or local patches are extracted from the videos, and feature vectors are constructed in the initial step. For achieving a robust descriptor against noise and positional uncertainties, 2D Gaussian blurring is applied in the second step. Dimensionality reduction for increasing processing speed is applied by random projection in the next step. These extracted features are classified with an ANN in the final step. This method is implemented on two different data sets. Experimental results show that by combining the last three steps and applying them after the first step (extracted features), efficiency is significantly improved compared to other methods. Recently 3D-CNN based architectures have attracted a great amount of attention for their automatic feature extraction property. They have various hyper parameters (layer numbers, filter numbers, kernel size, pool size, etc.) that must be tuned in an end-to-end manner. An improved 3D-CNN is designed as method 2 and its hyper parameters are adjusted and optimized. These two methods can be compared by their experimental results on two various types of data sets.

The major contributions are listed below:

1) A powerful traditional method for human activity recognition is proposed as method 1. It achieves significant efficiency compared to other methods.

2) An improved 3D-CNN is designed as method 2 for its popularity of deep-learning architectures.

3) A comparison of these two methods' performance is performed for various types of data sets. By investigating the efficiencies, it can be derived that method 1 gives remarkable efficiency by extracting powerful distinctive features while the efficiency of method 2 extremely relies on the diversity of data available for training.

The rest of this study is formed as follows: proposed methods are presented in the next section. Later the methods are implemented on two data sets, and experimental results are discussed. Finally, the performance of these two methods is investigated on different data sets, and their results are compared with those of the state-of-the-art approaches.

2- METHODS

two methods are introduced in this section. First, an ANN based architecture is proposed as method 1. Later, an extended 3D-CNN based architecture is presented as method 2.

2-1- Method 1: ANN based architecture

Method 1 consists of four steps as presented in Fig. 1: Interest point detection and features description, 2D Gaussian blurring (remove mean), dimensionality reduction (random projection) and classification using ANN, which are described below.

Interest points and local patches are extracted from the frames of video based on the local descriptors. Later, feature vectors are constructed. As mentioned in section [12], for achieving a powerful descriptor, 2D Gaussian blurring is performed in the spatial domain to ignore minor changes. It enhances the robustness of the descriptor against noise and positional uncertainties.

Achieved descriptors are high-dimensional and intensely

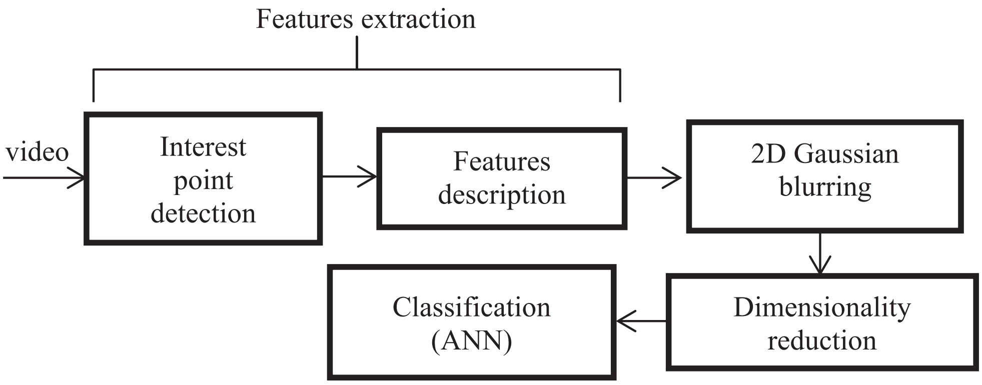


Fig. 1. Block diagram of method 1

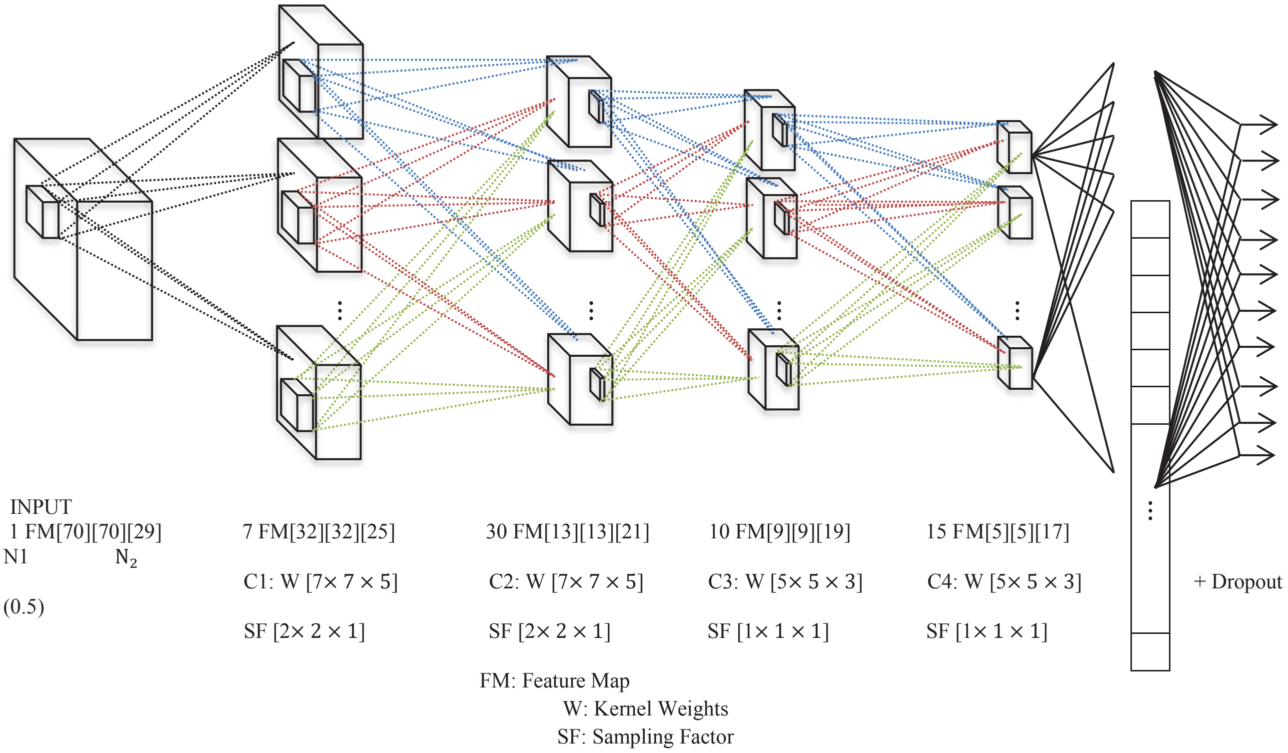


Fig. 2. The processing pipeline of the method 2

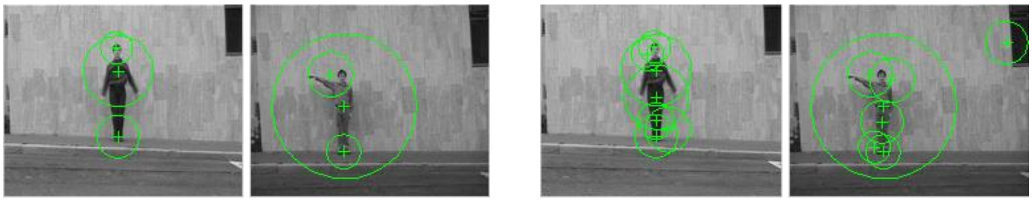
restricts the speed and applied applicability. A natural way is to decrease the dimensionality. Recently, random projection (RP) [21] has appeared as a strong tool in dimensionality reduction. Theoretical results demonstrate that the projections on a random lower dimensional subspace can retain the distances between vectors perfectly. The RP is data independence, simple, and fast [12].

Different studies had proven that artificial neural network can classify data accurately. It is the electronic network of neurons that is formed into layers called input, hidden and output. Multiple hidden layers can be in a neural network. It also contains one node for each class in the output layer. ANN generates its predictions in the form of a set of real values or Boolean decisions.

2-2- Method 2: 3D-CNN based architecture

To identify human action in videos along with spatial features, it is also essential to acquire the motion information encoded in several continuous frames. So 3D convolution is gotten by convolving a 3D kernel to the cube constructed by stacking several continuous frames together. By this formation, the feature-maps in the convolution layer are connected to several continuous frames in the previous layer, so acquiring motion information.

The improved layout of 3D-CNN designed for this task is shown in Fig. 2. In this layout, 29 consecutive frames with size of 70 × 70 are considered as input to the 3D convolutional neural network (input is [70 × 70 × 29] array). This design has seven layers containing the input layer. After the input layer,



a&b (3 points detect) c&d (7 points detect)

Fig. 3. Detected points in SURF descriptor. (a&b) 3 points detect, (c&d) 7 points detect

two convolution layers (C1 and C2) exist in which each layer has two steps (i.e. convolution followed by sub-sampling). They are followed by 3rd and 4rd convolution layers (C3 and C4), where no sub-sampling takes place (sub-sampling factor=1). They are later followed by neuron layer (N1 = 50), dropout with value of 0.5 to avoid overfitting and neuron layer (N2 = the number of actions), containing the output layer. The last two layers are fully-connected. First convolution layer (C1), contains 7 feature maps of size 32 × 32 × 25 (32 × 32 in spatial dimensions and 25 in temporal dimension). It is acquired by performing 7 different 3D kernels of size (7 × 7 × 5) tracked by sub-sampling of factor 2. Sub-sampling is done to resist small spatial distortions. Thus a CNN with multiple subsampling layers allows the processing of large inputs. Second convolution layer (C2), contains 30 feature maps of size 13 × 13 × 21. It is acquired by performing 3D kernels of size 7 × 7 × 5 tracked by sub-sampling of factor 2. Third convolution layer (C3), contains 10 feature maps of size 9 × 9 × 19 acquired by performing 3D kernels of size 5 × 5 × 3 on layer C2. Forth convolution layer (C4), contains 15 feature maps of size 5 × 5 × 17 acquired by performing 3D kernels of size 5 × 5 × 3 on layer C3. At this level, by applying several steps of convolution and subsampling, Spatio-temporal features are extracted from the input.

29 continuous input frames have been converted into a 6375D ((5 × 5 × 17) × 15) vector, taking the motion information in the input frames, after all the convolutional layers. Then the neuron layers (N1 and N2) act as a multilayer perceptron classifier on the 6375D input. N1 equals to 50 nodes and last layer has N2 nodes corresponding to N2 various actions. Back-propagation algorithm is applied for training this architecture.

3- Researches

In this section, researches were performed on two types of action data sets for comparing the performance of these two methods.

3-1- Data sets

Two public data sets were taken into consideration: Weizmann and KTH data sets. These data sets are very popular benchmark used in many action recognition studies.

Weizmann data set [22] includes 90 low-resolution videos of nine persons, each doing 10 actions: bend, jumping jack, jump forward, jump in place, run, gallop sideways, skip, walk, one-hand-waving and two-hands-waving. Their resolution is 180×144 pixels with 50 fps frame rate.

KTH data set [23] includes 599 videos with a length between 8 and 59 seconds. It comprises 6 kinds of human actions: walking, jogging, running, boxing, hand-waving and hand-clapping. Every activity is done into four various conditions. Videos were captured over similar backgrounds, using a fixed camera with 25 fps frame rate. Resolution is 160×120 pixels.

3-2- Experiment of method 1

The SIFT bundles a feature detector and descriptor. The detector extracts some related regions in a way, which is stable with some changes of the illumination, viewpoint and other viewing conditions. The descriptor attaches a sign to the sections, in order to strongly recognize their appearance. Each video splits into four frames and SIFT key points are detected for each frame. Descriptors are computed for each key point. The SIFT feature vector consists of 128 elements that are extracted from each key point. 2D Gaussian blurring is done on the SIFT feature vectors, and the random matrix used has dimensions of 100×128.

The 3D SIFT descriptor encodes the information of both space and time, and allows robustness for orientations and noise. Interest points are extracted from each frame of video. Thus, the 3D SIFT descriptors are constructed from these points, and consist of 640 elements. Furthermore, 2D Gaussian blurring and dimensionality reduction are done. Random matrix has 260×640 dimensions.

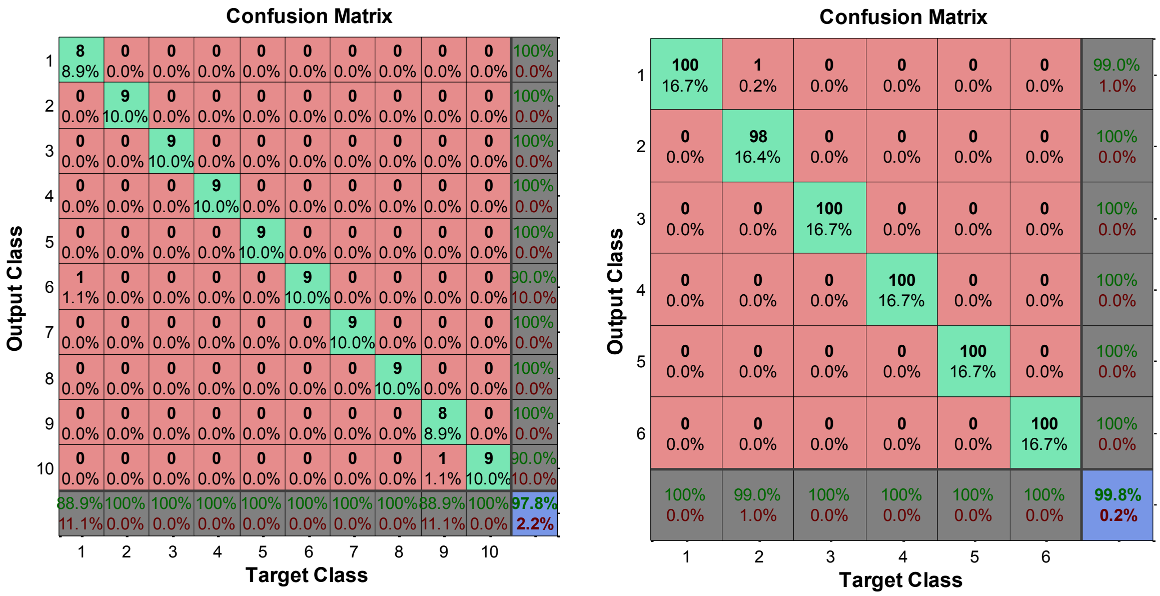
By using STIP (space-time interest point) [7], 40 of the strongest key points are extracted from each frame. Thus, for a video with N frames, we have (40 × N frames) matrix. In addition, 2D Gaussian blurring is performed here.

If the SURF algorithm is applied to the frames of each video, many points are detected to have no motion features. In addition, computation time increases and more memory is required. So the number of points are limited to 3. Different numbers of points that can be detected, are shown in Fig. 3. Each point has 64 features. These features are extracted from all the frames of video and 2D Gaussian blurring is perforemd.

For LMP [12], we set the spatial resolution (patch size) to 8 and the temporal resolution to 12. Descriptors computed from the spatiotemporal cubes points, P×D array, where P is the descriptor dimension (equals to 3×8×8 = 192). Second (variance), third (skewness) and fourth (kurtosis) central moments are calculated for each pixel along with the temporal direction. D is the total number of descriptors (i.e.

Table 1. Quantity evaluation of various descriptors on Weizmann data set

Local descriptors	Dimensions of input feature vector	Test time (ms)	Epoch	Accuracy (%)
SIFT(2D SIFT)	100	73.33	918	52.5
3D SIFT	260	4.63	64	95.2
STIP	40	5.46	406	87.1
SURF	64	16.01	363	93.8
LMP	192	38.66	949	79.0
HOG	22080	25.9	28	93.3
HOF	22080	27.48	29	94.4
HOG3D	**25000**	**31.54**	**87**	**97.8**



(a) (b)

Fig. 4. Confusion matrices resulting from ANN based architecture with HOG3D descriptor for (a) Weizmann,(b) KTH data sets

all key points extracted from 12 frames of each video). In addition,2D Gaussian blurring is done for each action.

Here, like section [11], descriptors were extracted on one scale where blocks contain 30×30 pixels ×30 frames, which at the same time is the dense sampling rate. Descriptors contain 2×2 spatial blocks and 2 temporal blocks. The magnitude responses to HOG, and HOF are divided into 8 orientations, achieving 64×345 (22080×1) dimensional descriptors for each video. Flow method, only used for HOF, is 'Horn-Schunck'.

By inspire HOG3D [5], 5 divisions in x and y spatial domains and 50 divisions in time domain, 25000 feature vectors were obtained for each video.

As described, input feature vectors were derived for the input of ANN. Output of ANN is a matrix with C × T dimension, which C is the number of classes and T is the second dimension of the input matrix. Each column has label, which detects the corresponding action. It should contain vectors of all zero values, except for a 1 in element I, where I is the class to show.

Finding the correct topology of ANN for a special usage is a challenging task and just through multiple implementations of various its topologies, in terms of numbers of neurons per layers and numbers of hidden layers, the optimized one can be detected. After multiple experiments, good results are achieved by applying a two-layer feed-forward network that has 75 and 50 neurons in the first and second hidden layers, in the Matlab environment. The patternnet() command is used to create a neural network "net", which can be trained for classifying inputs pursuant to target classes. The training function adopted was scaled conjugate gradient backpropagation ('trainscg'), which updates the weights and biases values of the ANN. The divide function is applied to divide the data into training, validation and testing subsets. The defined ratios for training, testing and validation are 0.7, 0.15 and 0.15, respectively. The performance function is set to cross entropy. Quantity evaluation of various descriptors on Weizmann data set is presented in Table 1. Achieved results show that ANN based architecture with HOG3D descriptor outperforms the others with 97.8% accuracy. To investigate the performance of this method, it is also performed on KTH data set, which gets 99.8% accuracy. Confusion matrices for these results are given in Fig. 4.

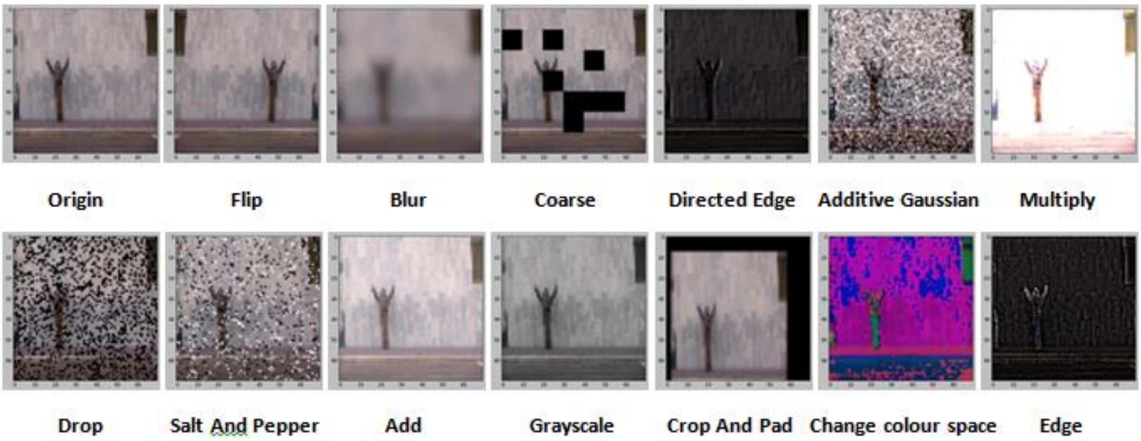


Fig. 5. Some transformations for augmentation on one frame of Weizmann data set

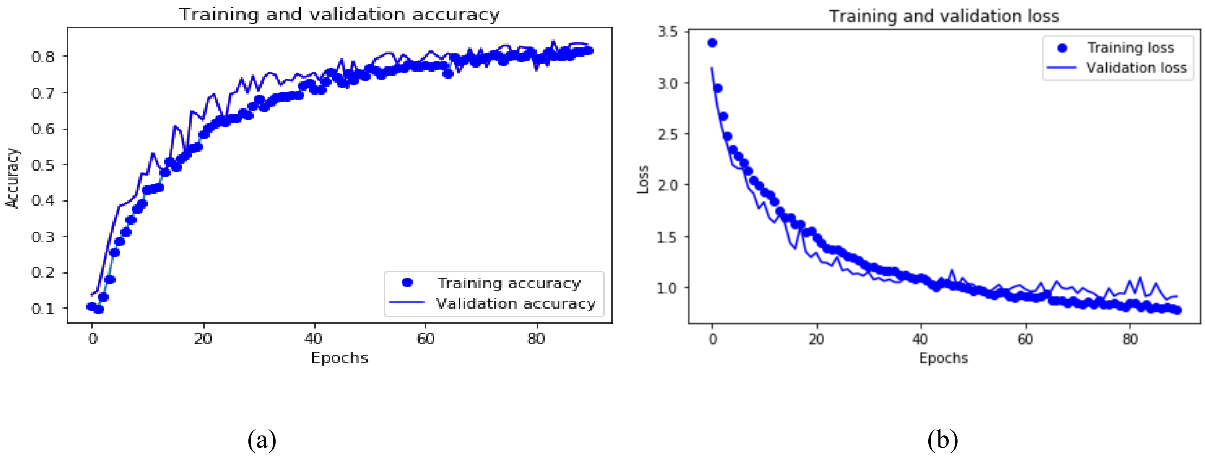


Fig. 6. (a) Accuracy and (b) Loss operations vs various epochs for Weizmann data set

3-3- Experiment of method 2

The extended 3D-CNN is trained and tested on these data sets. Since deep networks require to be trained on a huge number of training videos to get satisfactory efficiency, if the video consists of limited training videos, it is better to perform data augmentation to enhance the efficiency. There are many ways to do data augmentation, such as flipping, cropping, average blurring and so no. Therefore, data augmentation is done on training data sets. Some of these transformations on one frame of Weizmann data set are shown in Fig. 5.

The 3D-CNN model is trained over the training Weizmann data set for several epochs. It corresponds to a total of 406287 trainable parameters. The weights appeared to converge after 85 epochs, after which the accuracy became almost constant. The variation of accuracy and loss with number of epochs are depicted in Fig. 6. Fully trained 3D-CNN gives an accuracy of 83.3% when tested on Weizmann data set for 10 classifications. The test time for this method is 2.82 ms that is less than it for method 1. Four features which are automatically extracted from two frames with the first convolution layer are visualized in Fig. 7.

Subsequently, this 3D-CNN architecture corresponding to a total of 406083 trainable parameters is performed on KTH data set and achieves an accuracy of 92%. The confusion matrices that are resulted from this method are given in Fig. 8.

Results show that method 2 got better efficiency on KTH than Weizmann data set which is about 9%, though the same data augmentation was used for both of them. The reason is KTH data set consists of more original videos for training. This difference is only 2% for method 1, since it only relies on the quality of extracted features. In other words, the results of method 1 do not depend on the type of data sets, while the results of method 2 show this dependency. Thus, the efficiency of method 2 has been greatly reduced on Weizmann data set.

In addition, the performance of method 1 has also been compared with other methods, including Short-Time Motion Energy Image (ST-MEI) templates+Principal Component Analysis Network (PCANet) [3], improved hybrid feature extraction+Neural Network (NN) classifier [1], Scale Invariant Feature Transform (SIFT)+fine-tuning [24], and etc. From Table 2, we can see Short-Time Motion Energy Image (ST-MEI) templates+Principal Component Analysis Network (PCANet) [3] and improved hybrid feature extraction+Neural Network (NN) classifier [1] perform competitively with

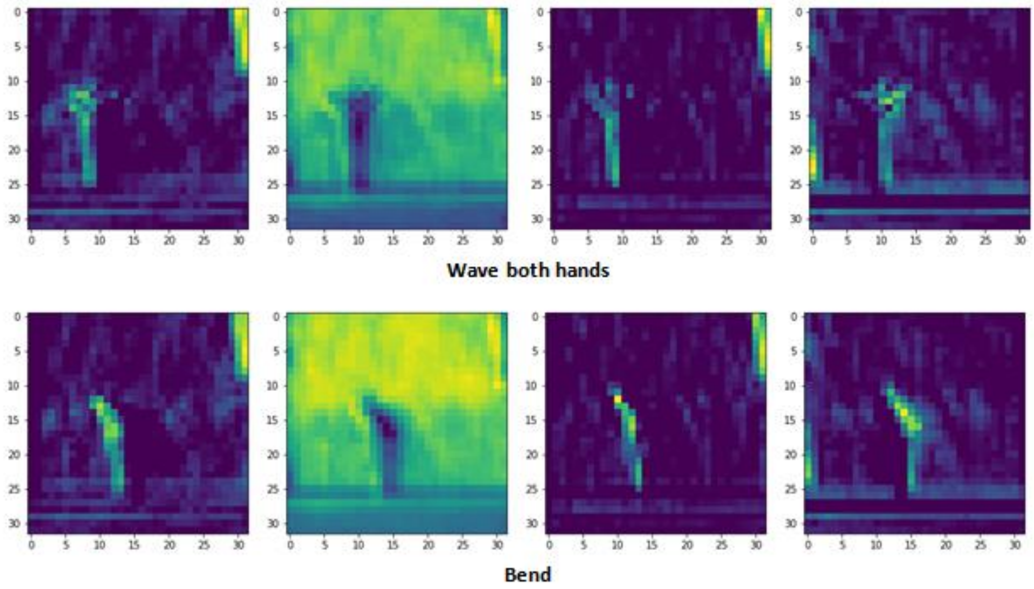
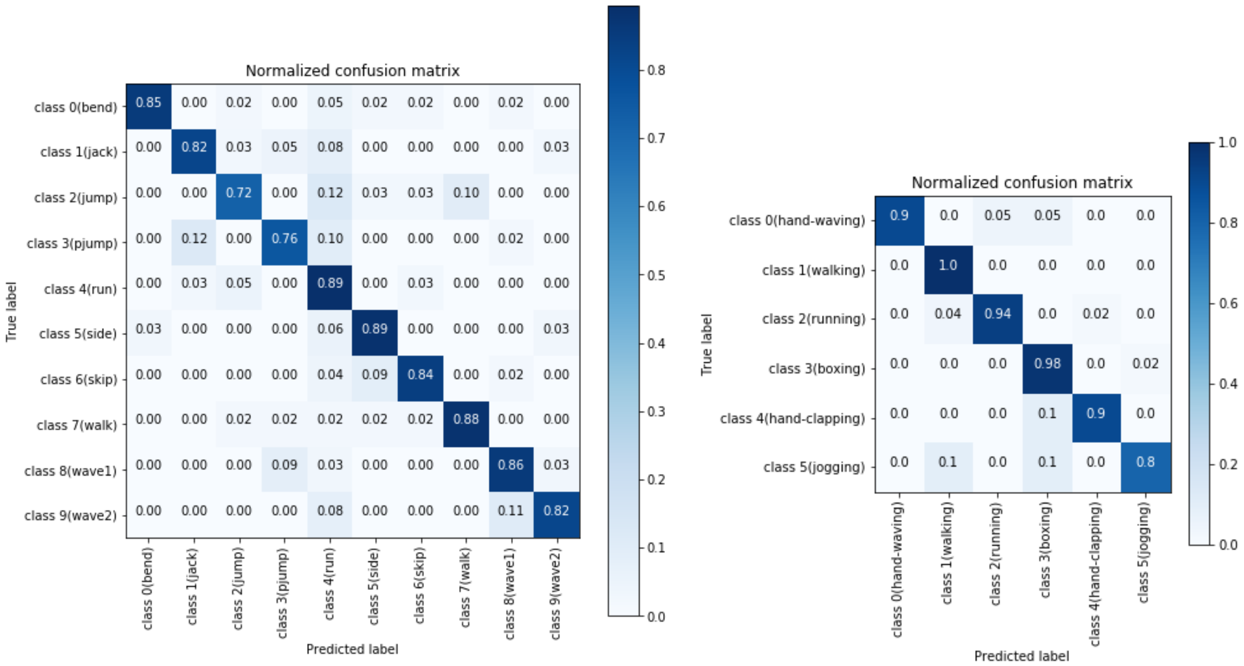


Fig. 7. Four features extracted from two frames with the first convolution layer (Upper row:Two-hands-waving,Lower row:Bend)



	class 0(bend)	class 1(jack)	class 2(jump)	class 3(pjump)	class 4(run)	class 5(side)	class 6(skip)	class 7(walk)	class 8(wave1)	class 9(wave2)
class 0(bend)	0.85	0.00	0.02	0.00	0.05	0.02	0.02	0.00	0.02	0.00
class 1(jack)	0.00	0.82	0.03	0.05	0.08	0.00	0.00	0.00	0.00	0.03
class 2(jump)	0.00	0.00	0.72	0.00	0.12	0.03	0.03	0.10	0.00	0.00
class 3(pjump)	0.00	0.12	0.00	0.76	0.10	0.00	0.00	0.00	0.02	0.00
class 4(run)	0.00	0.03	0.05	0.00	0.89	0.00	0.03	0.00	0.00	0.00
class 5(side)	0.03	0.00	0.00	0.00	0.06	0.89	0.00	0.00	0.00	0.03
class 6(skip)	0.00	0.00	0.00	0.00	0.04	0.09	0.84	0.00	0.02	0.00
class 7(walk)	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.88	0.00	0.00
class 8(wave1)	0.00	0.00	0.00	0.09	0.03	0.00	0.00	0.00	0.86	0.03
class 9(wave2)	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.11	0.82

(a)

	class 0(hand-waving)	class 1(walking)	class 2(running)	class 3(boxing)	class 4(hand-clapping)	class 5(jogging)
class 0(hand-waving)	0.9	0.0	0.05	0.05	0.0	0.0
class 1(walking)	0.0	1.0	0.0	0.0	0.0	0.0
class 2(running)	0.0	0.04	0.94	0.0	0.02	0.0
class 3(boxing)	0.0	0.0	0.0	0.98	0.0	0.02
class 4(hand-clapping)	0.0	0.0	0.0	0.1	0.9	0.0
class 5(jogging)	0.0	0.1	0.0	0.1	0.0	0.8

(b)

Fig. 8. Confusion matrices resulting from 3D-CNN based architecture for (a) Weizmann, (b) KTH data sets

method 1 for Weizmann and KTH data sets, respectively. However, it is necessary to note that one problem of Short-Time Motion Energy Image (ST-MEI) templates+Principal Component Analysis Network (PCANet) [3] is having several parameters that affect its recognition rate. The parameters of PCANet include filter size (patch size), the number of filters in each stage, the number of stages, the block size of the local histograms, and the ratio of overlapping between blocks. The optimal values for these parameters should be searched for each data set, which increases computational complexity. In improved hybrid feature extraction+Neural Network (NN) classifier [1], the choice of training/ testing ratio affects the classification accuracy. Varying the training/testing ratio from (60:40) to (70:30) for a Weizmann data set, enhances the classification accuracy from 97.20 to 100 because it does not have enough training/testing samples. While the recognition rate decreases for a KTH data set from 99.80 to 97.67 due to the availability of large amount of training data and clear chance of overfitting.

Table 2. Comparison accuracy of two proposed methods with existing human action techniques on both data sets

Weizmann data set

Reference	Year	Accuracy(%)
Scale Invariant Feature Transform (SIFT)+fine-tuning [24]	2015	96.66
Conditional Random Fields (CRFs) coupled with spatio-temporal Convolutional Neural Network (CNN) [25]	2016	93.30
view-invariant features+Nearest Mean Classifier (NMC) [2]	2018	95.56
improved hybrid feature extraction+Neural Network (NN) classifier [1]	2019	97.20
Discrete Wavelet Transform (DWT)+Support Vector Machine (SVM) classifier [4]	2020	97.00
Short-Time Motion Energy Image (ST-MEI) templates+Principal Component Analysis Network (PCANet) [3]	2020	100.0
method 1 (ANN based architecture)		97.80
method 2 (3D-CNN based architecture)		83.30

KTH data set

Reference	Year	Accuracy(%)
3D-Convolutional Neural Networks (3D-CNNs) [17]	2012	90.20
Scale Invariant Feature Transform (SIFT)+fine-tuning [24]	2015	97.89
Conditional Random Fields (CRFs) coupled with spatio-temporal Convolutional Neural Network (CNN) [25]	2016	95.50
view-invariant features+Nearest Mean Classifier (NMC) [2]	2018	90.58
developed 3D-CNN model [16]	2018	78.00
improved hybrid feature extraction+Neural Network (NN) classifier [1]	2019	99.80
encoding the distribution of video descriptors using Maclaurin coefficients of the density function and its moments [15]	2020	84.72
Short-Time Motion Energy Image (ST-MEI) templates+Principal Component Analysis Network (PCANet) [3]	2020	90.47
method 1 (ANN based architecture)		99.80
method 2 (3D-CNN based architecture)		92.00

4- CONCLUSION

In this paper, two methods are proposed for human activity recognition. In method 1, which is categorized into traditional approaches, different local descriptors have been used for features extraction. Later, these feature vectors are fed to artificial neural network for classifying. Among these descriptors, HOG3D achieves the best recognition accuracies of 97.8% and 99.8% for the Weizmann and KTH data sets, respectively.On the other hand, an improved 3D-CNN architecture is presented as method 2 for the same task. Experimental results show the dependency of these two methods on the data diversity. Thus, the accuracy of method 2 for Weizmann data set is significantly decreased compared to that for KTH data set, although data augmentation is done on training data to improve the model's performance and prevent overfitting.

In general, effective and distinctive features are required in traditional approaches, like method 1, whereas deep learning layouts, like method 2, need lots of original data to train for giving acceptable results. Indeed, learning Spatio-temporal features in deep learning layouts, such as method 2, is automatic while this step is time-consuming for traditional methods. Furthermore, it is important to note that in recent years, due to the popularity of deep learning layouts and limited volume of data, some hand-crafted features have also been added as inputs for training and improving their accuracy.

REFERENCES

[1] M.A. Khan, T. Akram, M. Sharif, M.Y. Javed, N. Muhammad, M. Yasmin, An implementation of optimized framework for action classification using multilayers neural network on selected fused features, Pattern Analysis and Applications, 22(4) (2019) 1377-1397.

[2] K.-P. Chou, M. Prasad, D. Wu, N. Sharma, D.-L. Li, Y.-F. Lin, M. Blumenstein, W.-C. Lin, C.-T. Lin, Robust feature-based automated multi-view human action recognition system, IEEE Access, 6 (2018) 15283-15296.

[3] A. Abdelbaky, S. Aly, Human action recognition using short-time motion energy template images and PCANet features, Neural Computing and Applications, (2020) 1-14.

[4] R. Singh, S. Nigam, A.K. Singh, M. Elhoseny, Intelligent Wavelet Based Techniques for Advanced Multimedia Applications, in, Springer, 2020.

[5] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in, 2008.

[6] R. Poppe, A survey on vision-based human action recognition, Image and vision computing, 28(6) (2010) 976-990.

[7] I. Laptev, On space-time interest points, International journal of computer vision, 64(2-3) (2005) 107-123.

[8] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision, 60(2) (2004) 91-110.

[9] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th ACM international conference on Multimedia, 2007, pp. 357-360.

[10] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Computer vision and image understanding, 110(3) (2008) 346-359.

[11] J. Uijlings, I.C. Duta, E. Sangineto, N. Sebe, Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off, International Journal of Multimedia Information Retrieval, 4(1) (2015) 33-44.

[12] T. Guha, R.K. Ward, Learning sparse representations for human action recognition, IEEE transactions on pattern analysis and machine intelligence, 34(8) (2011) 1576-1588.

[13] M.M. Moussa, E. Hamayed, M.B. Fayek, H.A. El Nemr, An enhanced method for human action recognition, Journal of advanced research, 6(2) (2015) 163-169.

[14] X. Sun, M. Chen, A. Hauptmann, Action recognition via local descriptors and holistic features, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2009, pp. 58-65.

[15] M. Saremi, F. Yaghmaee, Efficient encoding of video descriptor distribution for action recognition, Multimedia Tools and Applications, 79(9) (2020) 6025-6043.

[16] S.N. Boualia, N.E.B. Amara, 3D CNN for Human Action Recognition, 2018.

[17] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE transactions on pattern analysis and machine intelligence, 35(1) (2012) 221-231.

[18] V.A. Chenarlogh, F. Razzazi, Multi-stream 3D CNN structure for human action recognition trained by limited data, IET Computer Vision, 13(3) (2018) 338-344.

[19] G. Yu, T. Li, Recognition of human continuous action with 3D CNN, in: International Conference on Computer Vision Systems, Springer, 2017, pp. 314-322.

[20] K. Liu, W. Liu, C. Gan, M. Tan, H. Ma, T-c3d: Temporal convolutional 3d network for real-time action recognition, in: Thirty-second AAAI conference on artificial intelligence, 2018.

[21] R.G. Baraniuk, M.B. Wakin, Random projections of smooth manifolds, Foundations of computational mathematics, 9(1) (2009) 51-77.

[22] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, IEEE, 2005, pp. 1395-1402.

[23] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., IEEE, 2004, pp. 32-36.

[24] M.M. Moussa, E. Hamayed, M.B. Fayek, H.A. El Nemr, An enhanced method for human action recognition, Journal of advanced research, 6(2) (2015) 163-169.

[25] C. Liu, J. Liu, Z. He, Y. Zhai, Q. Hu, Y. Huang, Convolutional neural random fields for action recognition, Pattern recognition, 59 (2016) 213-224.

HOW TO CITE THIS ARTICLE

Zebhi, S., Almodarresi, S., Abootalebi, V. A Comparison of Two Neural Network Based Methods for Human Activity Recognition. AUT J. Elec. Eng., 53(1) (2021) 17-26.

DOI: 10.22060/eej.2020.18517.5357