Human Activity Recognition Based on Transfer Learning with Spatio-Temporal Representations

Saeedeh Zebhi, SMT Almodarresi, and Vahid Abootalebi
Electrical Engineering Department, Yazd University, Iran

Abstract: *A Gait History Image (GHI) is a spatial template that accumulates regions of motion into a single image in which moving pixels are brighter than others. A new descriptor named Time-Sliced Averaged Gradient Boundary Magnitude (TAGBM) is also designed to show the time variations of motion. The spatial and temporal information of each video can be condensed using these templates. Based on this opinion, a new method is proposed in this paper. Each video is split into N and M groups of consecutive frames, and the GHI and TAGBM are computed for each group, resulting spatial and temporal templates. Transfer learning with the fine-tuning technique has been used for classifying these templates. This proposed method achieves the recognition accuracies of 96.50%, 92.30% and 97.12% for KTH, UCF Sport and UCF-11 action datasets, respectively. Also it is compared with state-of-the-art approaches and the results show that the proposed method has the best performance.*

Keywords: *Deep learning, tuning, VGG-16, action recognition.*

Received August 5, 2020; accept January 6, 2021
https://doi.org/10.34028/iajit/18/6/11

1. Introduction

Human activity recognition is a great scene understanding ability for usages such as surveillance, robotics, and human-computer interactions. Activity recognition can be defined as the ability to recognize current activity on the basis of information received from a series of observations. Different approaches have been used to capture these activities. They can be categorized into vision-based and sensor-based [3]. In vision-based approaches, a camera is used to capture information about activities of human and visual features give basic information for actions. Then computer vision techniques can be applied for recognizing the human activities. In sensor-based approaches, different sensors are used to capture the behavior of human while they perform daily life activities. Vision-based approaches are pioneer approaches in this area and can provide good results. Therefore, they are focused in this literature.

As activity recognition has been an active research area newly, different methods have been proposed to deal with this task. They are separated into two main groups: Hand crafted features based methods and deep learning based methods. Also, several methods combine these two modalities. In hand crafted motion features methods, features are taken from the raw pixels of the video frames and are used to do the recognition. A lot of works use these hand crafted features and input them into the Support Vector Machine (SVM) classifier.

Along with advances in hand crafted motion features methods, deep neural networks have recently attracted much attention in areas such as object tracking, image segmentation and action localization, because the feature construction process is automated. In [8], features are extracted independently from each frame, then their predictions are pooled across the entire video. In fact, this approach completely ignores temporal structure. Also Two-Dimensional Convolutional Neural Networks (2D CNNs) are broadly used in image analysis applications, but they lose the capability to get motion information in terms of a 3D volume of video frames. Among 2D CNN methods, it is preferable to add a recurrent layer such as an LSTM to the model. It can code state and acquire temporal ordering. By adding it to the last layers of the 2D CNN model, high-level changes are modeled, but small low-level motion that is important in many cases may not be detected. It is also expensive to train because it needs unrolling the network through several frames for back-propagation-through-time [6].

Three-Dimensional (3D) convolutional neural network architecture is another usual deep models. By implementing 3D convolutions, features are extracted from both the spatial and the temporal dimensions, thereby acquiring the motion information, which is coded in several neighbor frames. It works better than single frame baseline since the motion features have been learned adequately. Because this architecture has more parameters than 2D CNN due to the extra kernel dimension, its training is more complex than 2D CNN.

As mentioned before, more 3D-CNN based approaches have high computational complexity. They concentrate on the problem of comprehending the content of videos, which does not essentially require

the modeling of their dynamics. Also the spatio-temporal filters of convolutional neural networks should be learned that maximize the recognition ability of the total system. On the other hand, the high efficiency of pre-trained networks in image classification problems encourage us to use them with little change for video. To achieve this purpose, we just need to construct some simple and informative templates which can effectively display the spatial and temporal information of videos. Additionally, these templates should not have high computational complexity. Gait History Image (GHI) [12] preserves the static (shape) information of sequences and it contains no temporal information. The usual static information is the proportion of the human body, clothing and objects, etc. An innovative descriptor called Time-Sliced Averaged Gradient Boundary Magnitude (TAGBM) is applied to detect the edges motion of objects which helps in tracking the objects in videos. Therefore, GHI and TAGBM are used for constructing templates that save and represent static and motion information of video, respectively.

The key contributions of this work can be abstracted as four fold:

1. The problem of human activity recognition is converted to templates classification;
2. For this purpose, GHI descriptor and the new designed descriptor named TAGBM are used for acquiring and abstracting spatial and temporal information of video to templates.
3. Combining these descriptors with deep learning architecture for human activity recognition is presented as another contribution.
4. An immense comparison with the state-of-the-art methods to accredit the performance of proposed method. Hence, a new practical method is proposed here. The spatial and temporal streams based on GHI and TAGBM are computed as features. Pre-trained network and fine-tuning techniques are used for the classification problem.

The rest of the paper is formed as follows: Related work is reviewed in part 2, Construction of GHI and TAGBM is explained in part 3, Proposed method is explained in part 4, Experiments are described in part 5, Several points about the method and results are discussed in part 6, and conclusion is presented in part 7.

2. Related Work

Human action recognition is becoming an interesting subject in computer vision. The ways of impressively presenting the spatial static and temporal dynamic information of videos are significant issues in this field. Ji *et al*. [7] proposed a 3D CNN model under uncontrolled environment. This model extracted features from both the spatial and the temporal dimensions by doing 3D convolutions. The features used for the final recognition task fused the attributes coming from multiple channels. Ramasinghe and Rodrigo [13] presented a CNN design, which gets motion and static information as inputs in a unit stream. This network can treat motion and static information as different feature maps and extract features off them, although stacked together. Zhou *et al*. [21] showed that a low dimension feature representation generated on the deep convolutional layers is more discriminative compared to traditional CNN features which explore the outputs from the fully connected layers in CNN.

Ullah *et al*. [17] proposed a method using CNN and deep bidirectional Long Short-Term Memory (LSTM) network. First, deep features are extracted from every sixth frame of the videos, which helps reduce the redundancy and complexity. Then, the sequential information among frame features is learnt using Densely-connected Bi-directional LSTM (DB-LSTM) network, where multiple layers are stacked together in both forward pass and backward pass of DB-LSTM. Wang *et al*. [18] proposed a new architecture that consists of CNN, LSTM units, and temporal-wise attention model. CNN is applied to extract spatial features and two kinds of LSTM networks are done on the spatial feature maps of various CNN layers to extract temporal features. After LSTM, a temporal-wise attention model is used to learn which parts in which frames are more significant. Finally, a joint optimization module is planned to explore intrinsic relations between two types of LSTM features. Wei *et al*. [19] proposed Pseudo-3D Convolutional Tube Network (P3D-CTN) to integrate the clip-based P3Dmodule with the frame-based 2D-module for achieving a balance between spatial independence and temporal continuity. The core part is the P3D-module, which gets the video as input and generates tube proposals associated with class scores. On the other hand, the 2D-module gets a single frame as input of 2D-CNN and gets tube proposals from P3D-module as Region of Interest (RoI). Ge *et al*. [5] proposed an attention mechanism based convolutional LSTM algorithm to enhance the efficiency of recognition by extracting the salient regions of actions efficiently. Zare *et al.* [20] proposed a Video Spatiotemporal Map (VSTM) of a video. VSTM is a compressed presentation of a video that combines its spatial and temporal features. It is generated by vertical concatenation of feature vectors produced from subsequent frames. VSTM enables CNNs to process a video for action recognition.

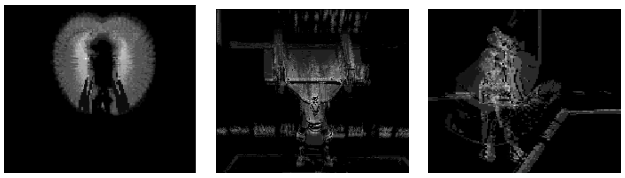
3. Descriptors

3.1. Gait History Image (GHI)

GHI is a specific information, which was introduced by Liu and Zheng [12]. It is constructed as follow:

$$E_{GHI}(x,y) = \begin{cases} P & if \quad B(x,y,n) = 1 \\ \sum_{n=1}^{p} D(x,y,n).(n-1) & otherwise \end{cases}, \quad D(x,y,n) = B(x,y,n+1) - B(x,y,n) \tag{1}$$

Where n is the frame number of image sequences, variables x and y are image pixels coordinates. $B(x,y,n)$ is acquired by BGSLibrary (Weighted Moving Variance) on the frame of video. By dividing all pixels values of $B(x,y,n)$ by the largest pixel value; that is 255, its pixel values are normalized. $D(x,y,n)$ is the binary difference image between two consecutive preprocessed binary sequences $B(x,y,n)$ and $B(x,y,n+1)$. $D(x,y,n)=1$ shows a motion occurrence in the th n time on the coordinate point (x,y). GHIs constructed for several consecutive frames are represented in Figure 1. As it is clear from this figure, moving pixels are brighter than other pixels. In fact, this template can show the spatial information of motion.



a) Hand-waving. b) Lifting. c) Tennis-serve.

Figure 1. GHIs for some consecutive frames.

3.2. Time-Sliced Averaged Gradient Boundary Magnitude (TAGBM)

We propose a novel informative template based on gradient images and Time-Sliced Averaged Motion History Image (TAMHI) [9], called TAGBM.

Suppose $F(x,y,n)$ is a denoised image sequence using median filtering. Here, n is the frame number of image sequences and variables x and y are image pixels coordinates. By applying simple 1-D [-1,0,1] Sobel masks on both x and y directions, image gradients are computed for it as follows:

$$F_x(x,y,n) = \frac{\partial F(x,y,n)}{\partial x} \quad , \quad F_y(x,y,n) = \frac{\partial F(x,y,n)}{\partial y} \tag{2}$$

Then, a [-1, 1] temporal filter is done over every two consecutive gradient sequences.

$$F_{n,x}(x,y,n) = \frac{\partial}{\partial n}\left(\frac{\partial F(x,y,n)}{\partial x}\right), \quad F_{n,y}(x,y,n) = \frac{\partial}{\partial n}\left(\frac{\partial F(x,y,n)}{\partial y}\right) \tag{3}$$

These time-derivatives of image gradients emphasize moving edge boundaries. Now the gradient boundary magnitude for each pixel (x,y) is defined as below:

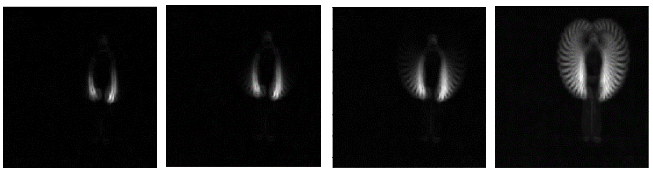
$$M(x,y,n) = \sqrt{F_{n,x}(x,y,n)^2 + F_{n,y}(x,y,n)^2} \tag{4}$$

For a sequence with p frames, TAGBM is

$$TAGBM = \frac{1}{\sum_{n=1}^{p} \alpha^{p-n}} \sum_{n=1}^{p} \alpha^{p-n} M(x,y,n) \tag{5}$$

Where α is the decay parameter ($0 < \alpha < 1$).

Choosing an appropriate α is important in characterizing the motion. The larger α is, the more detail of sequences is preserved. On the other hand, selecting a smaller α leads to loss of earlier trail of the edges motion. Figure 2 shows the TAGBMs at various values of the α for hand-waving.



a) α = 0.3. b) α = 0.5. c) α = 0.7. d) α = 0.9.

Figure 2. TAGBMs for some consecutive frames at various values of the α for hand-waving.

As seen, the temporal changes of the edges are more precise for α=0.9. So this value is selected to construct this template. Examples of TAGBM constructed for several consecutive frames are shown in Figure 3. It is obvious from this figure, more recent moving boundaries are brighter than other pixels. In fact, this template can show the time variation of edges motion.

a) Hand-waving. b) Lifting. c) Tennis-serve.

Figure 3. TAGBMs for some consecutive frames.

4. Proposed Method

4.1. Constructing Templates

First, each video splits into N groups of consecutive frames. GHIs are calculated for each group, creating GHIs of blocks. By using this approach, motion regions in N blocks of each video are abstracted in GHIs of blocks. It is clear that N has an important role. A small value of N causes that motion regions in the first frames of actions are confused with them in remaining frames; on the other side, by selecting a large value of N, fine motion regions in each group are considered which some of them may be produced because of background motions. Therefore, it is necessary to select the optimal value for N to maximize accuracy.

Similarity, TAGBMs are calculated for M groups of consecutive frames of each video, creating TAGBMs of blocks. GHIs include spatial information while TAGBMs contain motion information. So spatial and temporal streams are referred to them, respectively. These two individual streams are known as features and are separately fed to the classifier. Selecting the optimal value for M to maximize accuracy is also necessary in this stream. A small value of M causes the temporal differences of the first frames in actions to be ignored; on the other, by selecting a large value of M, fine temporal differences in each group are considered.

4.2. Classifications

Designing a convolutional neural network from scratch is complex and time-consuming, and training the weights from random values increases computational complexity. For classifying, transfer learning with fine-tuning technique is used. Two densely connected classifiers were added after the flatten layer on top of the convolutional base before compiling the model. The filter numbers for these two layers are 60 and C neurons, respectively, where C is the number of classes. To prevent overfitting, one dropout layer is added after the first dense layer. All Conv blocks of pre-trained network are frozen except Conv block 5 and the newly added top layers. The ReLU is used for the activation function. Classification is done with the softmax function and cross-entropy loss function.

This pre-trained network should have exactly three input channels, but GHIs and TAGBMs of blocks are greyscale images. The images simply need to appear to be Red, Green, And Blue (RGB). Therefore, the images are repeated three times on a new dimension. We will have them over all three channels, and the performance of this network should be the same as it was for RGB images.

Two probability matrices of size N-by-C and M-by-C are achieved with spatial and temporal streams for each video. By averaging the N+M row matrices, a predictive row matrix of size 1-by-C is generated. The column label which has the maximum value shows the predicted class. By using this proposed method, the problem of human activity recognition in video is converted to N+M templates classification and training of the network will be much easier and faster. The time complexity of computing N GHIs and M TAGBMs of blocks for each video is 0.61 and 4.95 seconds, respectively. This time is independent of the values of N and M. So proposed method optimizes time and computational complexity.

5. Experiments

Here three benchmark datasets were considered, which include the KTH [15], UCF Sport [16] and UCF-11 [11].

The KTH dataset 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 being done into four various conditions. Videos were captured over similar backgrounds with a fixed camera with 25 fps frame rate. Resolution of them is 160×120 pixels.

The UCF Sport dataset includes 150 video sequences with a resolution of 720×480. It comprises 10 actions that include walking, running, kicking, lifting, diving, golf swing, riding horse, skate boarding, swinging-side, and swinging-bench. These actions are performed in different real environments that cover various viewpoints and include a lot of camera motion.

The UCF-11 action dataset is a challenging dataset because of large variations in camera motion, illumination, viewpoint, background clutter, etc. It contains 1600 videos, each labeled as one of the following 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog.

Videos contain various number of frames. They are divided to N and M groups of non-overlapping consecutive frames. The GHIs and TAGBMs of blocks are calculated for these groups of frames, respectively. Considering N or M equal to 1 does not seem sensible to achieve high efficiency, the assumption that motion regions in the first frames are confused with them in residual frames. Also the movement changes that occur in the initial frames are ignored. RMSprop with a learning rate of 0.0001 is used for the optimization of all experiments. N and M start from 2 and increase until the best results are achieved. The batch size is fixed at 10 and the dropout rate is set to 0.8 in dropout layers to prevent overfitting. Also such as [11], five-fold cross-validation is used for experimental setups. Experiments have been implemented on a python based deep learning library called Keras with one NVIDIA GeForce GTX 1050 card and 8G RAM.

6. Results and Discussion

Each stream was initially performed to these datasets. The mean and standard deviation values for different N and M values are listed in Tables 1, 2, and 3. As shown, the best efficiencies are achieved with N=M=3 and N=M=4 for UCF Sport and UCF-11, respectively. For KTH, the best efficiencies are achieved with N=4 and M=3.

Table 1. Accuracy of KTH.

Training setting (N)	2	3	**4**	5
Spatial Stream	78.30±7.1%	80.81±8.5%	**93.65±1.9%**	92.99±2.55%
Training setting (M)	2	**3**	4	5
Temporal Stream	85.7±3%	**92.32±1.6%**	83.8±3.4%	72.6±2.1%

Table 2. Accuracy of UCF sport.

Training setting (N)	2	**3**	4
Spatial Stream	70±7.6%	**81.33±6.86%**	77.33±9.75%
Training setting (M)	2	**3**	4
Temporal Stream	75.33±6.86%	**81.33±2.7%**	78.3±5.12%

Table 3. Accuracy of UCF-11.

Training setting (N)	2	3	**4**	5
Spatial Stream	81.3±1.6%	84. 7 ±1.42%	**88.8±1.15%**	87.13±2.76%
Training setting (M)	2	3	**4**	5
Temporal Stream	80.5±3.49%	86.69±3.12%	**90.0±1.52%**	85.6±3.14%

The proposed method with the optimal N and M values for each dataset was applied, and the results are presented in Table 4. They show that spatial or temporal information is not sufficient for acquiring high accuracy. The results of proposed method show the effect of fusing these two streams. As shown, by considering two streams and fusing them, the accuracy is effectively improved. This performance improvement is obvious on datasets, as proposed method performs better than individual streams with an improvement of about 3.5% for KTH, 11% for UCF Sport and 7.7% for UCF-11.

Table 4. Results of individual streams and fusing them.

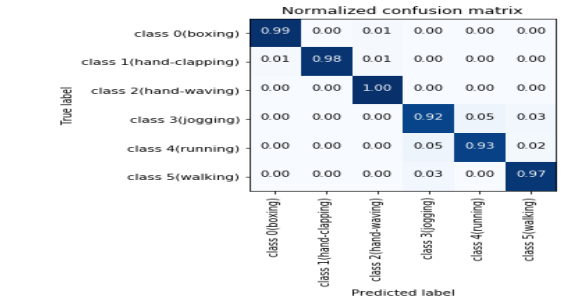
Datasets	Optimal N	Spatial Stream	Optimal M	Temporal Stream	Spatial+Temporal Stream
KTH	4	93.65±1.9%	3	92.32±1.6%	**96.50±0.63%**
UCF Sport	3	81.33±6.86%	3	81.33±2.7%	**92.3±0.02%**
UCF-11	4	88.8±1.15%	4	90.00±1.52%	**97.12±0.01%**

Experimental results show that most misclassification errors occur due to producing similar GHIs of blocks. It seems that by splitting each video to N groups of consecutive frames and changing the view angle of the camera throughout the videos, the GHIs of blocks acquired from some actions will be similar. In addition, the dynamic backgrounds increase this problem. This problem also exists for temporal stream because of producing homogeneous TAGBMs of blocks for actions with similar motions. By fusing the two streams in proposed method, some of these errors were corrected. Furthermore, often minor errors between different classes were corrected. Confusion matrices for the proposed method are shown in Figure 4.

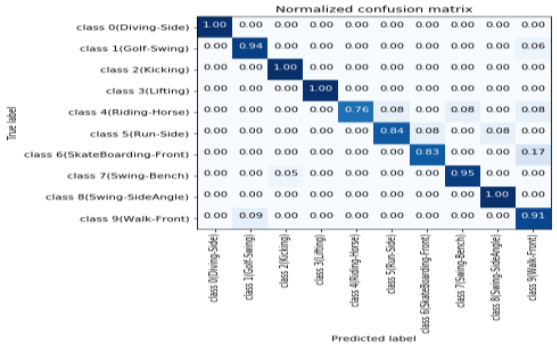
The proposed method had the best performance on these datasets. It was also compared to some previous methods in Table 5. As seen in this table, this method achieved better performance than others.

7. Conclusions

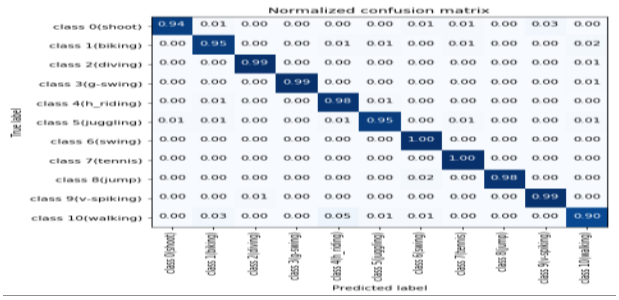
The main idea of this work is converting the problem of human activity recognition in video into templates classification. Descriptors like GHI and TAGBM are used for abstracting spatial and temporal information of video to templates. By combining the descriptors with deep learning architecture, a new method is proposed. First each video splits into N and M blocks, GHIs and TAGBMs are calculated for them. They extract spatial and temporal information of frames in video, respectively. Pre-trained network and fine-tuning techniques are used for classifying these templates. By using the proposed method, each video has abstracted to N+M templates, and there is no more need to save all frames of video. Less memory is needed, and the computational complexity is greatly reduced. Network training is also much easier and faster compared to 3D-CNN-based approaches. The results show that proposed method achieves significant recognition accuracy of 96.50%, 92.30% and 97.12% for KTH, UCF Sport and UCF-11 datasets, respectively. Then it is compared to state-of-the-art methods. In future work, we will attempt to amend these templates for confronting the interclass similarities such as dog-walking and horse-riding classes.



a) KTH.



b) UCF Sport.



c) UCF-11.

Figure 4. Confusion matrices resulting from proposed method for KTH, UCF Sport, UCF-11.

Table 5. Comparison of the proposed method with state-of-the-art methods.

Datasets	Methods	Accuracy
KTH	(Saremi and Yaghmaee)[14]	84.72%
	(Boualia and Amara) [2]	78.00%
	(Chou *et al.*,) [4]	90.58%
	(Liu *et al.*,) [10]	95.50%
	(Abdelbaky and Aly) [1]	90.47%
	Proposed method	**96.50%**
UCF Sport	(Saremi and Yaghmaee)[14]	70.67%
	(Zhou *et al.*,) [21]	90.00%
	(Wang *et al.*,) [18]	91.89%
	(Zare *et al.*,) [20]	82.14%
	(Wei *et al.*,) [19]	88.20%
	Proposed method	**92.30%**
UCF-11	(Ramasinghe and Rodrigo) [13]	93.10%
	(Ullah *et al.*,) [17]	92.80%
	(Wang *et al.*,) [18]	91.78%
	(Ge *et al.*,) [5]	94.12%
	Proposed method	**97.12%**

References

[1] Abdelbaky A. and Aly S., "Human Action Recognition using Short-Time Motion Energy Template Images and Pcanet Features," *Neural Computing and Applications*, vol. 23, pp. 12561-12574 , 2020.

[2] Boualia S. and Amara N., "3D CNN for Human Action Recognition," *in Proceedings of 18th International Multi-Conference on Systems, Signals and Devices*, Monastir, pp. 276-282, 2021.

[3] Chen L., Hoey J., Nugent C., Cook D., and Yu Z., "Sensor-Based Activity Recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790-808, 2012.

[4] Chou K., Prasad M., Wu D., Li D., Sharma N., Lin Y., Blumenstein M., Lin W., and Lin C., "Robust Feature-Based Automated Multi-View Human Action Recognition System," *IEEE Access*, vol. 6, pp. 15283-15296, 2018.

[5] Ge H., Yan Z., Yu W., and Sun L., "An Attention Mechanism based Convolutional LSTM Network for Video Action Recognition," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 20533-20556, 2019.

[6] Jaouedi N., Boujnah N., and Bouhlel M., "A Novel Recurrent Neural Networks Architecture for Behavior Analysis," *The International Arab Journal of Information Technology*, vol. 18, no. 2, pp. 133-139, 2021.

[7] Ji S., Xu W., Yang M., and Yu K., "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, 2012.

[8] Karpathy A., Toderici G., Shetty S., Leung T., Sukthankar R., and Fei-Fei L., "Large-Scale Video Classification with Convolutional Neural Networks," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, pp. 1725-1732, 2014.

[9] Lee C., Tan A., and Tan C., "Time-Sliced Averaged Motion History Image for Gait Recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 822-826, 2014.

[10] Liu C., Liu J., He Z., Zhai Y., Hu Q., and Huang Y., "Convolutional Neural Random Fields for Action Recognition," *Pattern Recognition*, vol. 59, pp. 213-224, 2016.

[11] Liu J., Luo J., and Shah M., "Recognizing Realistic Actions from Videos "in the wild"," *in Proceedings of Conference on Computer Vision and Pattern Recognition*, Miami, pp. 1996-2003, 2009.

[12] Liu J. and Zheng N., "Gait History Image: A Novel Temporal Template for Gait Recognition," *in Proceedings of IEEE International Conference on Multimedia and Expo*, Beijing, pp. 663-666, 2007.

[13] Ramasinghe S. and Rodrigo R., "Action Recognition by Single Stream Convolutional Neural Networks: An Approach Using Combined Motion and Static Information," *in Proceedings of 3rd IAPR Asian Conference on Pattern Recognition*, Kuala Lumpur, pp. 101-105, 2015.

[14] Saremi M. and Yaghmaee F., "Efficient Encoding of Video Descriptor Distribution for Action Recognition," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 6025-6043, 2020.

[15] Schuldt C., Laptev I., and Caputo B., "Recognizing Human Actions: A Local SVM Approach," *in Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, pp. 32-36, 2004.

[16] Soomro K. and Zamir A., *in Computer Vision in Sports*, Springer link, 2014.

[17] Ullah A., Ahmad J., Muhammad K., Sajjad M., and Baik S., "Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features," *IEEE Access*, vol. 6, pp. 1155-1166, 2017.

[18] Wang L., Xu Y., Cheng J., Xia H., Yin J., and Wu J., "Human Action Recognition by Learning Spatio-Temporal Features with Deep Neural Networks," *IEEE Access*, vol. 6, pp. 17913-17922, 2018.

[19] Wei J., Wang H., Yi Y., Li Q., and Huang D., "P3d-Ctn: Pseudo-3d Convolutional Tube Network for Spatio-Temporal Action Detection in Videos," *in Proceedings of IEEE International Conference on Image Processing*, Taipei, pp. 300-304, 2019.

[20] Zare A., Moghaddam H., and Sharifi A., "Video Spatiotemporal Mapping for Human Action Recognition by Convolutional Neural Network," *Pattern Analysis and Applications*, vol. 23, no. 1, pp. 265-279, 2020.

[21] Zhou Y., Pu N., Qian L., Wu S., and Xiao G., "Human Action Recognition In Videos of Realistic Scenes Based on Multi-Scale CNN Feature," *in Proceedings of Pacific Rim Conference on Multimedia*, Harbin, pp. 316-326, 2017.

Saeedeh Zebhi received the B.S. and M.S. degrees from the Department of Electrical Engineering, Yazd University of Iran, in 2009 and 2012, respectively. She is also currently a PHD candidate at yazd university. Her research interests include deep learning and video action recognition.

SMT Almodarresi obtained his B.S. degree in Electronics Engineering and M.S. degree in Communication Systems, both from the Isfahan University of Technology, Isfahan, Iran. He also holds Ph.D. in Electronics (Intelligent Signal Processing) from University of Southampton, UK (Department of Electrical and Computer Science: ECS). He works at the Department of Electrical and Computer Engineering in Yazd University where he pursues his research interests in: 1) Networked Control Systems (NCS) 2) Neuro-Fuzzy Networks 3) Wireless Networks.

Vahid Abootalebi received the B.S. and M.S. degrees in electrical Engineering from Sharif University of Technology, Tehran, Iran, in 1997 and 2000, respectively. He also received his Ph.D. degree in biomedical engineering from Amirkabir University of Technology, Tehran, Iran in 2006. Since 2007, he has been working as a faculty member of the Electrical Eng. Department of Yazd University, where he is currently an Associate Professor. His main research interests include biomedical signal processing and pattern recognition.