Human Action Recognition Using Transfer Learning with Spatio-Temporal Templates

S. Zebhi[1], SMT AlModarresi[*2], V. Abootalebi[3]

1- Electrical Engineering Department, Yazd University, Yazd, Iran, saeedehzebhi@gmail.com
2- Electrical Engineering Department, Yazd University, Yazd, Iran, smta@yazd.ac.ir
3- Electrical Engineering Department, Yazd University, Yazd, Iran, abootalebi@yazd.ac.ir

[*]Corresponding author,

Received: 2020-05-31

Revised: 2020-11-12

Accepted: 2020-12-10

Abstract

A gait energy image (GEI) is a spatial template that collapses regions of motion into a single image in which more moving pixels are brighter than others. The discrete wavelet transform template (DWT-TEMP) is a temporal template that represents the time changes of motion. The static and dynamic information of every video is compressed utilizing these templates. In the proposed method, every video is parted into N groups of successive frames, and the GEI and DWT-TEMP are made for every group, resulting spatial and temporal templates. Transfer learning method has been utilized for classifying. It gives the recognition accuracies of 92.40%, 95.30% and 87.06% for UCF Sport, UCF-11 and Olympic Sport action datasets, respectively.

Keywords

Discrete wavelet transform, gait energy image, human activity recognition.

1. Introduction

Recently human activity recognition (HAR) has attracted much attention in different usages like robotics, security, and health care. Different approaches can be utilized to record the activities which are split into vision-based and sensor-based [1]. Vision-based approaches are prior methods in this field and get acceptable results. So, they are concentrated in this research.

As HAR has been an important research area recently, various approaches [2-4] have been presented to solve this subject that are divided into two main groups: Hand crafted features based approaches and deep learning based approaches. Indeed, some approaches incorporate these two modalities. In the first group, appropriate features are extracted from the signals collected from sensors and are used for recognition task. Many works [5-10] utilize these selected features to train traditional machine learning (ML) models.

Along with advances in the first group, deep neural architectures [11-12] have absorbed much consideration in various research topics, since the feature construction process is automated.

GEI [13] saves static information of sequences and it has no temporal information. Pixels with more movement have higher intensity values. DWT is known to detect edges of an object that helps in tracking the object in videos. Its coefficients extracted local motion information of the object. Therefore, GEI and DWT are utilized for making templates that abstract and present static and moving information of video.

The main contributions of the paper are stated as two folds: 1) Descriptors like GEI, and discrete wavelet transform template (DWT-TEMP) are utilized for obtaining static and motion information of video and abstracting it to templates, so that the issue of HAR is changed to templates classification; and 2) combining the descriptors with deep model for HAR is presented as other contribution. So, a novel applied method is proposed here. By computing GEI and DWT-TEMP in two streams, feature vectors are constructed. Transfer learning technique is applied for classifying.

This research is expressed in several sections: Related work is summarized in section 2, Making of GEI and DWT-TEMP is described in section 3, Proposed approach is explained in section 4, Experiments are explained in section 5, Some important points of approach are described in section 6, and conclusion is represented in section 7.

2. Related Work

HAR is the most significant subjects in video processing. The methods of efficiently presenting the static and motion information of videos are considerable topics in this area. Ji et al. [14] proposed a three dimensional convolutional neural network architecture that static and motion features are extracted by applying three dimensional kernels. The features applied for the final recognition task fused the attributes coming from several channels. Ramasinghe and Rodrigo [15] introduced a convolutional neural network architecture, so that dynamic and static information is entered into the network in one stream and different features are extracted from information. Zhou et al. [16] presented that a low dimension feature presentation produced with the deep convolutional structure has more discriminative properties compared to common CNN features.

Ullah et al. [17] used CNN and bidirectional LSTM to propose a new approach. Features are extracted from each sixth frame of videos and then DB-LSTM network is applied for learning sequential information among them. Wang et al. [18] introduced a novel architecture that contains CNN, LSTM, and temporal-wise attention. Spatial features are extracted with CNN and then temporal features are extracted with two types of LSTMs. Finally, a temporal-wise attention is applied to recognize important sections of significant frames. A summary of the relevant work is presented in Table I.

Table I. An overview of the relevant work.

Approach	Author	Method
Hand-crafted feature	Efros et al. [25]	Optical flow based descriptor
	Shechtman and Irani [26]	Behavior-based similarity measure
	Hae Jong and Milanfar [27]	Local steering kernels computation as space-time descriptor
	Wang et al. [22]	Feature trajectories from dense optical flow
	Rahmani et al. [24]	Histogram of Oriented 3D Gradients (HOG3D) features and Locality-constrained Linear Coding (LLC)
Deep convolution feature	Ji et al. [14]	3D CNN model for extracting spatial and temporal features
	Ramasinghe and Rodrigo [15]	A CNN architecture with a single stream
	Zhou et al. [16]	Compared low dimension features extracted on the deep convolutional layers and traditional CNN features
	Ullah et al. [17]	Combined CNN and DB-LSTM
	Zhu et al. [23]	A deep LSTM network with the skeleton input

3. Descriptors

3.1. Gait Energy Image (GEI)

GEI is a specific information accumulation approach which was presented by Han and Bhanu [13] in 2004. It is constructed as follow:

$$GEI = \frac{1}{M}\sum_{n=1}^{M} B(x.y.n) \quad (1)$$

where n is the frame number of image sequences, variables x and y are image pixels coordinates. $B(x.y.n)$ is achieved by BackgroundSubtractorMOG2 on the frame of video. All pixel quantities of $B(x.y.n)$ are normalized by dividing them by maximum pixel quantity; it is 255. So, all regions of motion in a sequence are stacked in GEI.

Examples of GEI constructed for some successive frames are presented in Fig. 1. As it is obvious from this figure, more moving pixels are brighter than other pixels. This template presents the spatial information of motion.

3.2. Discrete Wavelet Transform Template (DWT-TEMP)

In this paper, DWT is used to compute the motion information. It is known to detect edges of an object, which helps in tracking the object in videos. A subset of wavelet coefficients is used to describe the movement of the object.

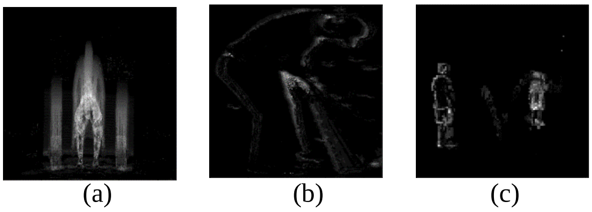
(a) (b) (c)

Fig. 1. GEIs for some successive frames. (a) lifting, (b) golf-swing, (c) shooting.

Each frame of video is decomposed into four sub-bands by computing the 2-D DWT and applying the haar wavelet. DWT gets approximation coefficients matrix cA and detail coefficients matrices cH, cV, and cD. Only three sub-bands of them have been used for constructing this template: cH, cV, cD. The template can be built by:

$$TEMP(n) = \frac{(cH + cV + cD)}{3}$$

$$DWT - TEMP = \frac{1}{M}\sum_{n=1}^{M} TEMP(n) \quad (2)$$

where n is the number of the frames in the image sequences. The expression $DWT - TEMP$ shows the motion information of the objects, so it is applied for action recognition in next sections. In other words, it shows the time information of object movement. DWT-TEMPs made for some successive frames are represented in Fig. 2.

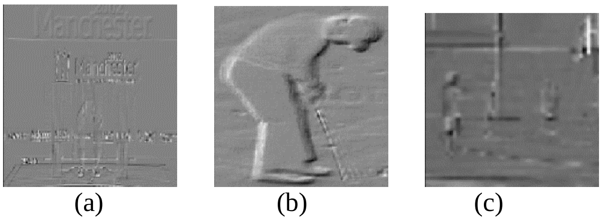
(a) (b) (c)

Fig. 2. DWT-TEMPs for some successive frames. (a) lifting, (b) golf-swing, (c) shooting.

4. Proposed Approach

4.1. Making templates

Every video parts into N groups of successive frames and GEIs are computed for them, producing GEIs of blocks. In this method, moving areas in N blocks of every video are summarized in GEIs of blocks. Obviously, N has a main effect. A tiny quantity of N induces that moving areas in the first frames of activities are distracted with them in remaining frames; Furthermore, by choosing a large quantity of N, tiny moving areas are remarked which are created due to background movements. So, choosing optimum quantity for N to enhance efficiency is essential.

Like before, DWT-TEMPs are computed for N groups of successive frames of every video, producing DWT-TEMPs of blocks. GEIs contain spatial information whereas DWT-TEMPs consist of motion information. Hence spatial and temporal streams are called to them. Feature vectors are constructed with two streams and then they are separately fed to the classifier as seen in Fig. 3. For achieving maximum efficiency, it is essential to choose optimum for N in this stream. By choosing a small quantity for N, the temporal differences of the first frames in activities are relinquished; also, by choosing a large quantity for N, tiny temporal changes in every group are remarked.

4.2. Classification

VGG-16 is a pretrained convolutional neural network with 16 layer. It was gotten in the best-performing image classification results. Also it is a very good architecture for benchmarking on a particular task and it is easy to implement. Two dense layers were appended on top of it which have 50 and C neurons (C is number of classes). All convolution blocks of this network are frosted except convolution block 5 and recently appended layers.

GEIs and DWT-TEMPs of blocks are greyscale templates. So, they are iterated three times on a novel dimension because pre-trained networks have been trained on RGB datasets.

Two probability matrices of size N -by- C are obtained with two streams. Rows of these matrices are averaged, so that a predictive row matrix is produced. The column label with maximum quantity exhibits predicted class. In this way, the problem of HAR in video is transformed to 2*N templates classification. The time complexity of calculating N DWT-TEMPs and N GEIs of blocks for a video with 100 frames is 0.1 and 0.3 seconds, respectively.

5. Experiments

Three popular datasets were used, which contain UCF Sport [28], UCF-11 [29] and Olympic Sport [30]. Videos include different number of frames which are parted to N groups of successive frames. The GEIs and DWT-TEMPs of blocks are computed for them, respectively. Assuming N equal to 1 does not be reasonable to get high accuracy, the presumption that moving areas in the first frames are distracted with them in remaining frames. Furthermore, moving variations that happen in the first frames are ignored. Root mean square propagation function is applied for optimizing all experiments. N starts from 2 and increases till the best results are acquired. Indeed, five-fold cross-validation is utilized for experiments.

6. Results

First, every stream was done. The mean and standard deviation quantities for various N quantities are shown in Table II to IV. The dependent sample t-test is also done for comparing the means between two related groups of datasets. As presented, the best accuracies are acquired with N = 3, N = 4 and N = 5 for UCF Sport, UCF-11 and Olympic Sport, respectively. For UCF Sport, all p-values are less than 0.05. It is also true for UCF-11 except for N = 3, 5 in two streams. In these cases, the p-values are larger than 0.05 that means a considerable difference does not exist between N =3 and N=5. For Olympic Sport, all p-values are less than 0.05 except for N = 4, 6 in temporal stream. So in this case, the null hypothesis is rejected and there's no difference between the means of these two groups.

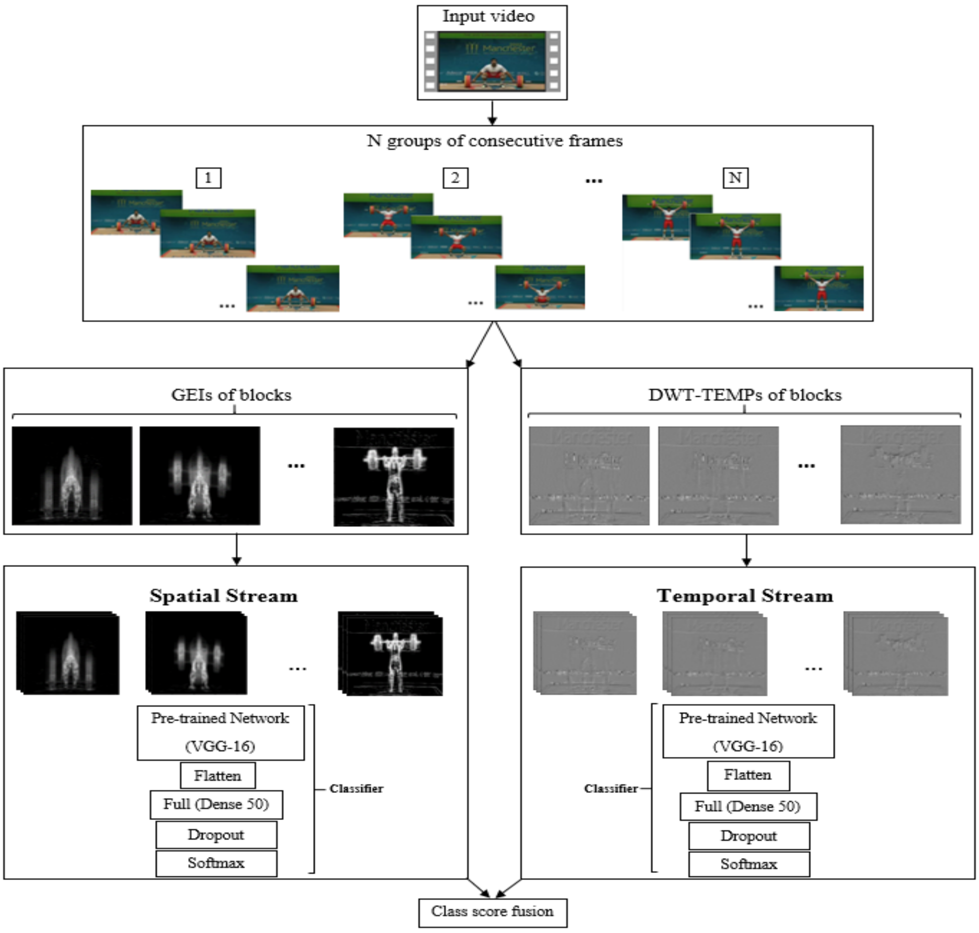

Fig. 3. Flowgraph of proposed method.

Table II. (a) Accuracy of UCF Sport, (b) Comparison between two groups.

(a)

Training setting (N)	2	**3**	4
Spatial Stream	64±7.1%	**75.3±1.2%**	72.6±2%
Temporal Stream	75.7±3%	**85.7±2.1%**	81.3±0.8%

(b)

Comparison (N)	2, 3	2, 4	3, 4
		p-value	
Spatial Stream	0.019	0.036	0.005
Temporal Stream	0.008	0.024	0.016

Table III. (a) Accuracy of UCF-11, (b) Comparison between two groups.

(a)

Training setting (N)	2	3	**4**	5
Spatial Stream	72.7±1%	74±0.6%	**79.5±0.8%**	74.1±1.4%
Temporal Stream	85.5±1.1%	87.2±0.6%	**89.9±0.3%**	87.5±1.2%

(b)

Comparison (N)	2, 3	2, 4	2, 5	3, 4	3, 5	4, 5
			p-value			
Spatial Stream	0.024	0.001	0.015	0.001	**0.849**	0.002

Temporal Stream	0.043	0.001	0.001	0	**0.730**	0.009

Table IV. (a) Accuracy of Olympic Sport, (b) Comparison between two groups.

(a)

Training setting (N)	2	3	4	**5**	6
Spatial Stream	67.4±1.2%	71.3±1.1%	74.5±0.8%	**78.1±1%**	73.3±0.6%
Temporal Stream	72±1.4%	74.1±0.7%	78.6±0.9%	**81.6±1.1%**	77.8±0.7%

(b)

Comparison (N)	2, 3	2, 4	2, 5	2, 6	3, 4	3, 5	3, 6	4, 5	4, 6	5, 6
	p-value									
Spatial Stream	0.001	0	0	0	0.006	0	0.014	0.002	0.005	0
Temporal Strean	0.027	0.003	0	0.002	0.001	0	0.001	0.011	**0.133**	0.007

The results of proposed approach with optimal N quantities for three datasets are presented in Table V.

It can be concluded that each of these streams is not enough for achieving desirable efficiency. The affect of fusing these two streams can be understood from the results of proposed approach, so that the accuracies are efficiently improved. It is clear that the proposed approach works better compared to each stream as shown in Table VI. Further, the time complexity of the proposed approach is showed in this Table. For temporal stream, it is 3 times smaller than it for spatial stream. For the proposed approach, it is equal to the sum of them for each stream. Also, it is independent of N value.

6.1. Limitations

Results present that most errors happen because of generating analogous GEIs of blocks. In fact, some reasons like parting every video to N groups of successive frames, varying angle of camera and dynamic backgrounds cause to produce these errors. Similarity, analogous DWT-TEMPs of blocks are produced in temporal stream for above reasons.

The two streams are fused in proposed approach, so some of the mention errors were corrected. Also, often small errors between various classes were corrected. The proposed approach had the best accuracy on datasets. In addition, the efficiency of proposed approach has been compared with others, containing Multi-scale CNN feature [16], Spatiotemporal features with DNNs [18], Single stream CNN [15], Attention mechanism based Conv- LSTM [20], Motion hierarchies [3], ProtoGAN [4], etc. As seen in Table VII, the proposed approach gived better accuracies compared to other approaches on three datasets. Of course, Spatiotemporal features with DNNs [18], Attention mechanism based Conv- LSTM [20] and ProtoGAN [4] do competitively with the proposed approach for these datasets.

7. Conclusion

Transforming the issue of HAR in video into templates classification is the principal purpose of this research. Descriptors like GEI and DWT-TEMP are applied for saving spatial and temporal information of video. In a proposed approach, they are combined with deep structure. GEIs and DWT-TEMPs are computed for every video by splitting it into N blocks. They extract various information of frames in video. Transfer learning technique is utilized for classifying these templates. By applying the proposed approach, 2*N templates are acquired from every video, therefore saving total frames of video is not necessary. Lesser memory is required; also computational complication is highly decreased. Network training is too simpler and quicker contrasted to three-dimensional convolutional neural network methods. The proposed approach has been tested on three well-known datasets and it gets significant human activity recognition accuracy for them.

Datasets	Optimal N	Fold1	Fold2	Fold3	Fold4	Fold5	Avg±SD
UCF Sport	3	93.33	96.67	90	90	93.33	**92.4±2.1%**
UCF-11	4	94.38	94.69	96.25	95.31	95.63	**95.3±0.7%**
Olympic Sport	5	87. 9	87.26	87.26	87.18	85.26	**87.06±0.8%**

Table V. Results of proposed approach.

Table VI. Results of individual streams and fusing them.

Datasets	N	Spatial Stream	Temporal Stream	**Spatial+Temporal Stream**	**Time complexity**
UCF Sport	3	75.3±1.2%	85.7±2.1%	**92.4±2.1%**	**38.34 sec**
UCF-11	4	79.5±0.8%	89.9±0.3%	**95.3±0.7%**	**1224.24 sec**
Olympic Sport	5	78.1±1%	81.6±1.1%	**87.06±0.8%**	**727.16 sec**

Table VII. Comparison.

Datasets	Methods	Accuracy
UCF Sport	Visual saliency (Souly & Shah., 2016) [7]	85.10%
	Encoding of video descriptor (Saremi et al., 2020) [8]	70.67%
	Multi-scale CNN feature (Zhou et al., 2017) [16]	90.00%
	Spatiotemporal features with DNNs (Wang et al., 2018) [18]	91.89%
	P3d-ctn (Wei et al., 2019) [19]	88.20%
	Proposed method	**92.40%**
UCF-11	Visual attention (Sharma et al., 2015) [6]	84.90%
	Single stream CNN (Ramasinghe et al., 2015) [15]	93.10%
	Bi-directional LSTM (Ullah et al., 2017) [17]	92.80%
	Spatiotemporal features with DNNs (Wang et al., 2018) [18]	91.78%
	Attention mechanism based Conv- LSTM (Ge et al., 2019) [20]	94.12%
	Proposed method	**95.30%**
Olympic Sport	Motion hierarchies (Gaidon et al., 2014) [3]	85.50%
	ProtoGAN (Dwivedi et al., 2019) [4]	86.30%
	Multi-model feature fusion (Cong et al., 2020) [10]	72.36%
	3D CNN (Ji et al., 2012) [14]	68.20%
	Proposed method	**87.06%**

References

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

[2] H.-H. Phan, N.-S. Vu, V.-L. Nguyen, and M. Quoy, "Action recognition based on motion of oriented magnitude patterns and feature selection," *IET Computer Vision*, vol. 12, no. 5, pp. 735-743, 2018.

[3] A. Gaidon, Z. Harchaoui, and C. Schmid, "Activity representation with motion hierarchies," *International journal of computer vision*, vol. 107, no. 3, pp. 219-238, 2014.

[4] S. K. Dwivedi, V. Gupta, R. Mitra, S. Ahmed, and A. Jain, "ProtoGAN: Towards Few Shot Learning for Action Recognition," *arXiv preprint arXiv*:1909.07945, 2019.

[5] J. Cho, M. Lee, H. J. Chang, and S. Oh, "Robust action recognition using local motion and group sparsity," *Pattern Recognition*, vol. 47, no. 5, pp. 1813-1825, 2014.

[6] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv*:1511.04119, 2015.

[7] N. Souly and M. Shah, "Visual saliency detection using group lasso regularization in videos of natural scenes," *International Journal of Computer Vision*, vol. 117, no. 1, pp. 93-110, 2016.

[8] M. Saremi and F. Yaghmaee, "Efficient encoding of video descriptor distribution for action recognition," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 6025-6043, 2020.

[9] Y. Zhang, M. Ding, Y. Bai, D. Liu, and B. Ghanem, "Learning a strong detector for action localization in videos," *Pattern Recognition Letters*, vol. 128, pp. 407-413, 2019.

[10] J. Cong and B. Zhang, "Multi-model feature fusion for human action recognition towards sport sceneries," *Signal Processing: Image Communication*, p. 115803, 2020.

[11] S. Javanmardi, A. Latif, and V. Derhami, "Image Tag Completion by Applying SPFCM Clustering on the Features Learned by Deep Convolutional Neural Networks," *TABRIZ JOURNAL OF ELECTRICAL*

ENGINEERING, vol. 49, no. 1, pp. 111-123, 2019.

[12] A. Sezavar, H. Farsi, and S. Mohamadzadeh, "Content-Based Image Retrieval using Deep Convolutional Neural Networks," *TABRIZ JOURNAL OF ELECTRICAL ENGINEERING*, vol. 48, no. 4, pp. 1595-1603, 2019.

[13] J. Han and B. Bhanu, "Statistical feature fusion for gait-based human recognition," *in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004. CVPR 2004., 2004, vol. 2: IEEE, pp. II-II.

[14] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221-231, 2012.

[15] S. Ramasinghe and R. Rodrigo, "Action recognition by single stream convolutional neural networks: An approach using combined motion and static information," *in 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR),* 2015: IEEE, pp. 101-105, 2015.

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

[17] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155-1166, 2017.

[18] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," *IEEE access*, vol. 6, pp. 17913-17922, 2018.

[19] J. Wei, H. Wang, Y. Yi, Q. Li, and D. Huang, "P3d-ctn: Pseudo-3d convolutional tube network for spatio-temporal action detection in videos," *in 2019 IEEE International Conference on Image Processing (ICIP),* 2019: IEEE, pp. 300-304.

[20] H. Ge, Z. Yan, W. Yu, and L. Sun, "An attention mechanism based convolutional LSTM network for video action recognition," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 20533-20556, 2019.

[21] A. Zare, H. A. Moghaddam, and A. Sharifi, "Video spatiotemporal mapping for human action recognition by convolutional neural network," *Pattern Analysis and Applications*, vol. 23, no. 1, pp. 265-279, 2020.

[22] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," *in CVPR IEEE 2011*, pp. 3169-3176, , 2011.

[23] W. Zhu et al., "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," *arXiv preprint arXiv*:1603.07772, 2016.

[24] H. Rahmani, D. Q. Huynh, A. Mahmood, and A. Mian, "Discriminative human action classification using locality-constrained linear coding," *Pattern recognition letters*, vol. 72, pp. 62-71, 2016.

[25] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," *in null, 2003*, p. 726: IEEE.

[26] E. Shechtman and M. Irani, "Space-time behavior based correlation," *in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05),* 2005, vol. 1, pp. 405-412: IEEE.

[27] H. J. Seo and P. Milanfar, "Action recognition from one example," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 867-882, 2010.

[28] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," *in Computer vision in sports*: Springer, 2014, pp. 181-208.

[29] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos “in the wild”," *in 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009: IEEE, pp. 1996-2003.

[30] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," *in European conference on computer vision*, 2010: Springer, pp. 392-405.