Three-Stream Network With Bidirectional Self-Attention for Action Recognition in Extreme Low Resolution Videos

Didik Purwanto, Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang

***Abstract*—This letter presents a novel three-stream network for action recognition in extreme low resolution (LR) videos. In contrast to the existing networks, the new network uses the trajectory-spatial network, which is robust against visual distortion, instead of the pose information to complement the two-stream network. Also, the new three-stream network is combined with the inflated 3D ConvNet (I3D) model pre-trained on kinetics to produce more discriminative spatio-temporal features in blurred LR videos. Moreover, a bidirectional self-attention network is aggregated with the three-stream network to further manifest various temporal dependence among the spatio-temporal features. A new fusion strategy is devised as well to integrate the information from the three different modalities. Simulations show that the new architecture outperforms the main state-of-the-art extreme LR action recognition methods on the HMDB-51 and IXMAS datasets.**

***Index Terms*—Action recognition, low resolution videos, self-attention, trajectory-spatial network, deep learning.**

I. Introduction

LOW resolution (LR) videos arise in a variety of disciplines such as video surveillance [1]–[3], action recognition [4]–[8], and face detection [9]–[11]. However, LR videos in general contain less visual information and are susceptible to noise. It is thus challenging to develop a robust descriptor for action recognition in LR videos.

A myriad of algorithms has been addressed for action recognition in extreme LR videos. Ryoo *et al.* [1] introduced inverse super resolution, which takes advantage of the existing high resolution videos in training by learning different types of sub-pixel transformations. Chen *et al.* [12] proposed a semi-coupled network, which is based on filter sharing to benefit from high resolution training. Rahman *et al.* [13] combined the handcrafted and the deep learned features to improve performance. Ryoo *et al.* [14] used a two-stream multi-siamese convolutional neural network (CNN) to learn shared embedding spaces that map LR videos with the same content to the same location. Also, Yu *et al.* [15] proposed a pseudo tensor low rank regularization to recover inherent robust components of an input video. Xu *et al.* [16] proposed a fully-coupled network architecture to generate robust video representation by incorporating 3D Convolutional and RNN to better capture motion information. However, the aforementioned methods [1], [12]–[16] did not fully exploit the temporal relationships among frames, which is beneficial in learning action recognition when there is a substantial loss of spatial information. Some recent approaches such as 3D skeletal [17], [18] or differential images [19] were also considered for action recognition, but they were not devised for extreme LR videos.

In this letter, we present a novel three-stream network for action recognition in extreme LR videos. To resolve the visual degradation, in contrast to the existing ones [20], [21], our three-stream network uses trajectory patterns in the Hue, Saturation, Value (HSV) color space instead of the pose information to encode trajectory temporal dynamic information, the former of which is more robust against visual distortion, to complement the well-known two-stream networks. Also, the new network is combined with the inflated 3D ConvNet (I3D) model [22] pre-trained on Kinetics to produce more discriminative spatio-temporal features in blurred LR videos. Moreover, a bidirectional self-attention network is aggregated with the three-stream network to further manifest various temporal dependency among spatio-temporal features. A new fusion strategy is devised as well to integrate the information from the three different modalities. Simulations show that the new approach provides superior performance over the state-of-the-art extreme LR action recognition methods on the HMDB-51 and IXMAS datasets.

The contributions of this letter can be summarized as follows: i) we employ the trajectory-spatial information to capture the fine-grained motion in extreme LR videos, which can complement the conventional two-stream network; ii) we propose a new architecture, which combines the three-stream network with a bidirectional self-attention network based on a new pairwise similarity function to leverage the temporal dependency information; (iii) we design a new fusion strategy to effectively aggregate the outputs from the three different modalities; iv) we demonstrate that the I3D model pre-trained on a large-scale video dataset such as Kinetics can benefit action classification in extreme LR videos.

Manuscript received February 27, 2019; revised May 23, 2019; accepted June 6, 2019. Date of publication June 19, 2019; date of current version July 2, 2019. This work was supported by the Ministry of Science and Technology, China under contracts MOST 107-2221-E-011-124 and MOST 107-2221-E-011-078-MY2. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yap-Peng Tan. (*Corresponding author: Didik Purwanto.*)

The authors are with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan (e-mail: d10602806@mail.ntust.edu.tw; d10702801@mail.ntust.edu.tw; ytchen@mail.ntust.edu.tw; whf@mail.ntust.edu.tw).

Digital Object Identifier 10.1109/LSP.2019.2923918

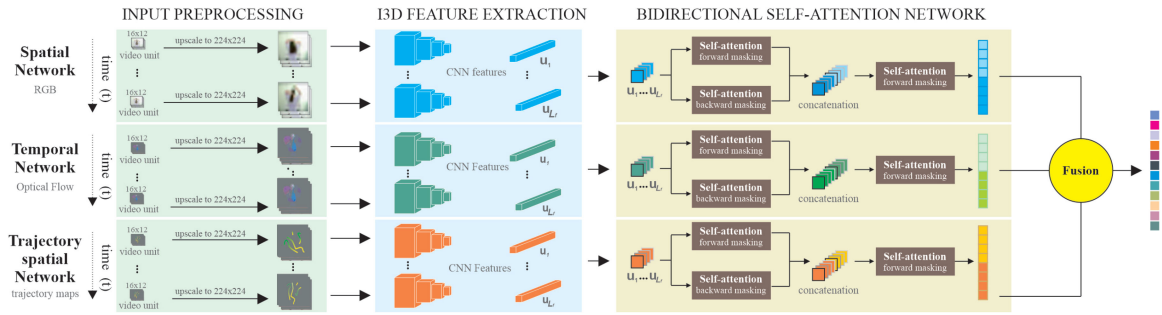

Fig. 1. An overall pipeline of the proposed architecture that utilizes a three-stream network with the I3D model as the backbone to extract a sequence of spatio-temporal features and a bidirectional self-attention network to leverage temporal relationships, followed by a new fusion scheme to attain the final classification scores.

II. Proposed Method

A. Three-Stream Network

Due to a significant loss of spatial information, it is difficult to extract foreground action movement from LR videos. To overcome this setback, we consider a new three-stream network, which comprises of a two-stream network, spatial and temporal CNNs, and a trajectory-spatial network to capture the trajectory information. The input to the trajectory-spatial network is the newly generated trajectory-spatial images based on the improved dense trajectory [23] to capture the subtle movement within homogeneous spatial areas in the HSV color space. The trajectory-spatial images aim to encode the history of the trajectory patterns over the time. Given a video and a set of K trajectories $\{D_k^m\}_{k=1}^K$ within a temporal window, beginning at frame m, with a fixed length T_l, the HSV color coding of a trajectory $D_{k_0}^m$ that occupies spatial locations $\{d_1^{k_0,m}, \ldots, d_{T_l}^{k_0,m}\}$ is given by

$$\text{hue}(k_0, m) = \frac{m}{F} \frac{\sum_{t_l=1}^{T_l-1} |d_{t_l+1}^{k_0,m} - d_{t_l}^{k_0,m}|}{\max_{k=1}^{K} \sum_{t_l=1}^{T_l-1} |d_{t_l+1}^{k,m} - d_{t_l}^{k,m}|} \quad (1)$$

$$\text{saturation}(t_l) = \alpha \frac{t_l}{T_l}, \quad t_l = 1, \ldots, T_l \quad (2)$$

where α is a hyperparameter that controls the smoothness of the saturation and F denotes the number of frames in the video. Based on (1) and (2), the motion speed and the temporal position information can be encoded by the hue and the saturation channels, respectively.

B. Spatio-Temporal Feature Extraction

The I3D feature extraction, inflating the 2D convolutional filters into the 3D counterparts, is combined with the three-stream network described above to produce the deep spatio-temporal descriptors. Now, the spatial and the temporal CNNs process a stack of consecutive RGB frames and optical flow images, respectively. We fine-tune each stream independently using the cross-entropy loss function and high resolution pre-trained model.

Following [16], we upscale 16×12 LR images into 224×224 as the input to the I3D model. We then partition the input video into T_f non-overlapping video units $V_1, \ldots, V_{T_f}$, each of which contains $L = F/T_f$ frames. Thereafter, we extract spatio-temporal features from each video unit using the last convolutional layer. The feature dimension for every video unit is $N \times C$, where $N < L$) is the pooled temporal length and C is the number of the feature channels.

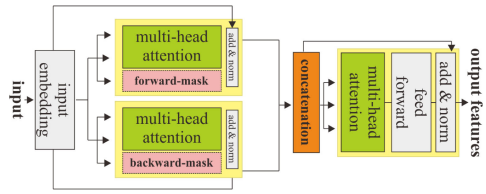

Fig. 2. The bidirectional self-attention network.

C. Bidirectional Self-Attention Network

For a sequence of video units of length $L_f = N \times T_f$, we employ a bidirectional self-attention network, which is an effective technique for natural language inference and sentiment analysis [24], [25], to capture various temporal relationships among the spatio-temporal features. The bidirectional self-attention network, as shown in Fig. 2, is a modification of transformer encoder [26], in which position encoding is invoked based on a pairwise function to generate more representative temporal features from a sequence of video units.

The core module of the transformer encoder is called multi-head attention, where several soft-attention layers run in parallel and their outputs are concatenated. Self-attention mechanism computes non-local response at a position in a sequence by relating information from all other positions. Given spatio-temporal features $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_{L_f}] \in \mathbb{R}^{L_f \times C}$ from a sequence of video units, a non-local function in deep neural networks at position i, $\mathbf{n}_i$, is defined as [26]:

$$\mathbf{n}_i = \mathbf{u}_i + \frac{1}{b(\mathbf{u}_i)} \sum_j f(\mathbf{u}_i, \mathbf{u}_j) g(\mathbf{u}_i) \quad (3)$$

where $b(\cdot)$ is a normalization factor and $g(\cdot)$ is a linear embedding layer. The pairwise function $f(\mathbf{u}_i, \mathbf{u}_j)$ computes the similarity between the features at positions i and j and is given by

$$f(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{u}_i^T \mathbf{W}_i \mathbf{u}_j + e_1(\mathbf{u}_i)^T e_2(\mathbf{u}_j) \quad (4)$$

where $e_1(\cdot)$ and $e_2(\cdot)$ are linear embedding layers and $\mathbf{W}_i \in \mathbb{R}^{C\times C}$ is a linear weight matrix for the position i. As with [26], we divide the feature channel of dimension C into H heads and apply (3) to each head. We then concatenate the output of the non-local operator on every head to obtain a feature representation $\mathbf{N} = [\mathbf{N}^1, \mathbf{N}^2, \ldots, \mathbf{N}^H]$ with the same dimension as the original sequence input, where $\mathbf{N}^h = [\mathbf{n}_1^h, \ldots, \mathbf{n}_{L_f}^h] \in \mathbb{R}^{L_f\times(C/H)}$, $h = 1, \ldots, H$.

In this letter, we modify the previous pairwise function in (4) in two ways. First, to capture the periodic patterns of the actions, we consider the discrete fourier transform (DFT) [27] of the feature vectors. Second, since the pairwise function in (3) is symmetric, it does not exhibit temporal positional information. Inspired by [24], [25], we add in bidirectional self-attention to facilitate the encoding of the temporal order information. Toward this end, we first perform masking in both of the forward and backward time directions on the non-local function using either a positional forward mask $\mathcal{M}^{fw}$ or a backward mask $\mathcal{M}^{bw}$ as follows [26]:

$$\begin{cases} \mathcal{M}_{i,j}^{fw} = 0, & i < j \\ -\infty, & \text{otherwise} \end{cases} \tag{5}$$

$$\begin{cases} \mathcal{M}_{i,j}^{bw} = 0, & i > j \\ -\infty, & \text{otherwise.} \end{cases} \tag{6}$$

Consequently, by including the positional masks, the new pairwise function $\bar{f}(\mathbf{u}_i, \mathbf{u}_j)$ can be expressed as

$$\bar{f}(\mathbf{u}_i, \mathbf{u}_j) = \sigma(\mathcal{M}_{i,j}^{d} + \tilde{f}(\mathbf{u}_i, \mathbf{u}_j)) \tag{7}$$

where $\mathcal{M}_{i,j}^{d}$ is either $\mathcal{M}_{i,j}^{fw}$ or $\mathcal{M}_{i,j}^{bw}$, depending on the direction of the positional mask, $\sigma(\cdot)$ is a sigmoid function [28] that normalizes the features, and

$$\tilde{f}(\mathbf{u}_i, \mathbf{u}_j) = \tilde{\mathbf{u}}_i^T \mathbf{W}_i \tilde{\mathbf{u}}_j + e_1(\mathbf{u}_i)^T e_2(\mathbf{u}_j) \tag{8}$$

in which $\tilde{\mathbf{u}}_i$ is the DFT of $\mathbf{u}_i$.

The outputs from the bidirectional self-attention network are concatenated and passed on to the self-attention mechanism without any positional mask to obtain the final feature representation. This network is trained by using the cross-entropy loss function.

D. New Fusion Strategy

This subsection considers a new fusion method to efficaciously combine the features from the three different modalities. Since the features from each stream have a different significance, we thus apply different weights to each stream to produce a more faithful final feature representation. Such an aggregation can minimize the overfitting problem when multiple modalities are directly combined in the late fusion stage [29].

Denote the features from the spatial, the flow, and the trajectory-spatial streams as $\mathbf{N}_s$, $\mathbf{N}_f$, and $\mathbf{N}_t$, respectively, where $\mathbf{N}_s, \mathbf{N}_f, \mathbf{N}_t \in \mathbb{R}^{L_f\times 2C}$. The final feature representation, $\bar{\mathbf{N}}$, is then determined by

$$\bar{\mathbf{N}} = (\mathbf{w}_s \mathbf{1}^T) \odot \mathbf{N}_s + (\mathbf{w}_f \mathbf{1}^T) \odot \mathbf{N}_f + (\mathbf{w}_t \mathbf{1}^T) \odot \mathbf{N_t} \tag{9}$$

TABLE I
PARAMETER SETTINGS FOR TRAINING THE THREE-STREAM AND THE BIDIRECTIONAL SELF-ATTENTION NETWORKS

	Three-stream	Attention modules
Pre-trained model	Kinetics+ImageNet	-
Video unit length	64	-
Optimizer	Adam	Adam
Learning rate	0.0001	0.00001
Epoch	20	70

TABLE II
THE PERFORMANCE OF THE PROPOSED ARCHITECTURE WITH VARIOUS MECHANISMS

Spatial	Temporal	Trajectory-spatial	Bidirectional self-attention	Fusion	Accuracy	
					HMDB-51	IXMAS
✓	-		-	-	37.30	90.56
	✓		-	-	48.75	90.80
	-	✓	-	-	22.03	73.88
✓	✓	-	-	-	52.61	93.89
✓	✓	✓	-	-	52.81	94.23
✓	✓	✓	✓	-	54.60	97.00
✓	✓	✓	✓	✓	56.93	97.78

where $\odot$ is element-wise multiplication [30] and $\mathbf{1} \in \mathbb{R}^{\mathbf{2C}}$ is a vector of all ones. The weights $\mathbf{w}_s, \mathbf{w}_f, \mathbf{w}_t \in \mathbb{R}^{L_f}$ are adjusted in the same way, e.g., $\mathbf{w}_s$ is determined by

$$\mathbf{w}_s = \sigma\left(\mathbf{N}_s \mathbf{b}_k + \frac{(\mathbf{N}_s + \mathbf{N}_f + \mathbf{N}_t)}{3} \mathbf{w}_k\right) \tag{10}$$

where $\mathbf{b}_k, \mathbf{w}_k \in \mathbb{R}^{2C}$ are linear weights learned during the network training process. We use the fully connected and softmax layers to generate action class scores.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Low Resolution Datasets

To create LR videos, the original videos in the HMDB-51 [31] and IXMAS [32] datasets are downsampled into 16×12 resolution using the average downsampling [14]. The HMDB-51 dataset consists of 51 classes with the resolution being 320×240. This dataset poses several difficulties such as occlusion, background clutter, and different viewpoints. The IXMAS dataset, recorded with different actors, cameras, and viewpoints, comprises of 12 classes with the resolution being 64×48.

B. Experimental Setup and Evaluation Protocol

The learning process is comprised of training the three-stream network and the bidirectional self-attention module, which are conducted separately. The hyperparameters of these training procedures are listed in Table I for reference. The simulations mainly follow the protocols and evaluation metrics provided by [16], [31] for HMDB-51 and [12], [32] for IXMAS.

C. Ablation Studies

We assess the performance of the I3D embedded two-stream network with and without the trajectory-spatial network, bidirectional self-attention, and the new fusion on the LR HMDB-51 and IXMAS datasets, as shown in Table II, from which we can see that the trajectory-spatial network entails lower accuracy compared with the other streams because it is trained from the scratch. The trajectory-spatial information can enhance

TABLE III
COMPARISON OF THE ACTION RECOGNITION RESULTS ON THE LOW RESOLUTION HMDB-51 AND IXMAS DATASETS

Method	Accuracy	
	HMDB51	IXMAS
pLRN+Tennet [15]	21.70	-
ISR [1]	28.68	-
Semi-Coupled [12]	29.20	93.70
Rahman *et al.* [13]	33.74	-
C3D [8]	35.16	-
C3D [8] + trajectory-spatial stream	35.48	-
Multi-Siamese [14]	37.70	-
Fully-Coupled [16]	44.96	-
I3D [22]	52.61	93.89
I3D [22] + self-attention [26]	53.55	95.00
Ours	**56.93**	**97.78**

the performance of the two-stream network by 0.2% to 0.3%. Also, combining bidirectional self-attention can boost the performance by about 1.8% to 2.8% due to its capability to model the temporal relationships among videos. Finally, together with the new fusion, the performance can be further improved by about 0.8% to 2.4%, as it can generate more faithful action class scores.

D. Comparison With the State-of-the-Art Methods

In this subsection, we compare the proposed architecture with some state-of-the-art methods, pLRN+Tennet [15], ISR [1], Semi-Coupled [12], Rahmat *et al.* [13], Multi-Siamese [14], C3D [8], Fully-Coupled [16], and I3D [22] on the LR HMDB-51 dataset, where [8], [16] and [22] used a pre-trained model from the Sport-1M and Kinetics datasets, respectively, both of which are large-scale datasets, and [1], [12]–[14] are all trained on the ImageNet dataset. From Table III we can see that pLRN+Tennet [15] produces the worst performance as it considered the low-rank video representation instead of the more effective features by CNN. ISR [1] provides better performance by incorporating inverse super resolution algorithm to learn different types of sub-pixel transformations. Semi-Coupled [12] is slightly better than ISR by utilizing filter sharing in the two-stream networks. Also, [13] excels previous methods by leveraging the context information and optical flow information. The combination of C3D [8], which is based on 3D ConvNet, and our trajectory-spatial network provides higher accuracy. Multi-Siamese [14] further improves the performance by applying max pooling to different time intervals. We can also find that Fully-Coupled [16] is superior to the aforementioned approaches by using temporal alignment information through the integration of C3D and gated recurrent units. I3D [22] attains the second best performance, as it extracts the deep learned features from multiple frames to produce more discriminative spatio-temporal features. The performance of I3D is further boosted by incorporating the self-attention mechanism [26] to learn the temporal dependency. Our new three-stream architecture achieves the best performance by taking advantage of the transfer learning from Kinetics, including the trajectory-spatial information to explore the fine-grained motion movement, and employing bidirectional self-attention to model the temporal dependency.

We also make comparisons on the LR IXMAS dataset, as shown in Table III, in which I3D that learns multiple frame information at once can yield better accuracy than Semi-Coupled [12]. I3D is also boosted by learning the temporal dependency using the self-attention mechanism [26]. Our three-stream approach again outperforms the other three baselines by further exploiting the fine-grained motion movement and modelling various temporal dependency using bidirectional self-attention.

IV. CONCLUSION

This letter has developed an efficacious three-steam network, comprising of the spatial, temporal, and trajectory-spatial networks, bundled with the I3D model for spatial-temporal feature extraction and a bidirectional self-attention network to capture various temporal dependency. A new fusion scheme is also addressed to aggregate the information from different modalities. Simulations show that the new architecture surpasses the state-of-the-art methods on the extreme LR HMDB-51 and IXMAS datasets.

REFERENCES

[1] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *Proc. Assoc. Advancement Artif. Intell.*, 2017, pp. 4255–4262.
[2] H. K. Chen, X. G. Zhao, S. Y. Sun, and M. Tan, "PLS-CCA heterogeneous features fusion-based low-resolution human detection method for outdoor video surveillance," *Int. J. Autom. Comput.*, vol. 14, no. 2, pp. 136–146, 2017.
[3] M. Haghighat and M. Abdel-Mottaleb, "Low resolution face recognition in surveillance systems using discriminant correlation analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 912–917.
[4] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2008, pp. 1–8.
[5] Z. Gao, G. Lu, and P. Yan, "Enhancing action recognition in low-resolution videos using Dempster-Shafer's model," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, 2016, pp. 676–680.
[6] Y. Zhao, H. Di, J. Zhang, Y. Lu, F. Lv, and Y. Li, "Region-based mixture models for human action recognition in low-resolution videos," *Neurocomputing*, vol. 247, pp. 1–15, 2017.
[7] Y. Yang, R. Liu, C. Deng, and X. Gao, "Multi-task human action recognition via exploring super-category," *Signal Process.*, vol. 124, pp. 36–44, 2016.
[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 4489–4497.
[9] W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 327–340, Jan. 2012.
[10] S. Shekhar, V. M. Patel, and R. Chellappa, "Synthesis-based robust low resolution face recognition," 2017, arXiv:1707.02733.
[11] T. Yoshida, T. Takahashi, D. Deguchi, I. Ide, and H. Murase, "Robust face super-resolution using free-form deformations for low-quality surveillance video," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2012, pp. 368–373.
[12] J. Chen, J. Wu, J. Konrad, and P. Ishwar, "Semi-coupled two-stream fusion ConvNets for action recognition at extremely low resolutions," in *Proc. IEEE Winter Conf. Appl. Comput. Vision*, 2017, pp. 139–147.
[13] S. Rahman, J. See, and C. C. Ho, "Deep CNN object features for improved action recognition in low quality videos," *Adv. Sci. Lett.*, vol. 23, no. 11, pp. 11360–11364, 2017.
[14] M. S. Ryoo, K. Kim, and H. J. Yang, "Extreme low resolution activity recognition with multi-siamese embedding learning," in *Proc. Assoc. Advancement Artif. Intell.*, 2018, pp. 7315–7322.
[15] T. Yu, L. Wang, C. Guo, H. Gu, S. Xiang, and C. Pan, "Pseudo low rank video representation," *Pattern Recognit.*, vol. 85, pp. 50–59, 2019.
[16] M. Xu, A. Sharghi, X. Chen, and D. J. Crandall, "Fully-coupled two-stream spatiotemporal networks for extremely low resolution action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vision*, 2018, pp. 1607–1615.

[17] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao, "Discriminative multi-instance multitask learning for 3D action recognition," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 519–529, Mar. 2017.

[18] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, and X. Gao, "Latent max-margin multitask learning with skelets for 3-D action recognition," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 439–448, Feb. 2017.

[19] D. Xie, C. Deng, H. Wang, C. Li, and D. Tao, "Semantic adversarial network with multi-scale pyramid attention for video classification," in *Proc. Assoc. Advancement Artif. Intell.*, 2019, pp. 2866–2869.

[20] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2923–2932.

[21] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "Potion: Pose motion representation for action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7024–7033.

[22] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6299–6308.

[23] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[24] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, arXiv:1810.04805.

[25] T. Shen, T. Zhou, G. Long, J. Jiang, and C. Zhang, "Bi-directional block self-attention for fast and memory-efficient sequence modeling," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–18.

[26] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[27] H. F. M. Zaki, F. Shafait, and A. Mian, "Modeling sub-event dynamics in first-person action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 7253–7262.

[28] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 2625–2634.

[29] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L. P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2016, pp. 284–288.

[30] W. McGuire, R. H. Gallagher, and R. D. Ziemian, *Matrix Structural Analysis*. Hoboken, NJ, USA: Wiley, 2000.

[31] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 2556–2563.

[32] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 635–648.