IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 1, February, 2023, pp. 185-190
Journal homepage: http://iieta.org/journals/ria

Action Recognition Using Segmental Action Network

Hakim Nasaoui*, Insaf.Bellamine, Hassan Silkan

LAROSERI, Department of computer science, Chouaïb Doukkali University Faculty of Sciences, El Jadida 24000, Morocco

Corresponding Author Email: HakimNasaoui@gmail.com

https://doi.org/10.18280/ria.370123

Received: 5 January 2023
Accepted: 10 February 2023

Keywords:
3D convolutional neural network, STIP, deep learning, Optical flow

ABSTRACT

Human action recognition refers to the task of recognizing and categorizing human actions in video or image sequences. This is a complex problem in computer vision and has a wide range of applications, including video surveillance, human-computer interaction, and sports analysis. In this work, A novel approach to action recognition using Segmental Action Networks is presented. The proposed approach utilizes 2D and 3D convolutional neural networks to extract spatiotemporal features from video frames, which are used to train a Segmental Action Network. To improve the model's accuracy, Different voting and feature extraction techniques, such as Space-Time Interest Points, (STIP) and Optical flow have been applied. The proposed model has been tested on the HMDB51 dataset and has achieved better results than existing models. The results demonstrate the effectiveness and robustness of our proposed approach for action recognition. Furthermore, our model is computationally efficient and can be deployed on edge devices with low computational and memory capacity, making it a promising approach for real-world applications.

1. INTRODUCTION

Human action recognition is a task in computer vision that aims to recognize and classify human actions in videos or images. It is used in applications such as video surveillance, human-computer interaction [1], and video indexing. The task is challenging due to the complexity of human movement and the lack of labeled data. Different methods have been proposed to address the task, including discriminative learning and deep learning approaches. Recent advances in deep learning have enabled significant progress in this field.

Deep neural networks have been used to great effect in human action recognition due to their ability to extract meaningful features from raw data and to learn complex patterns. This enables them to learn patterns, such as motions in a video, that would otherwise be difficult to capture using traditional methods. As a result, deep neural networks are capable of accurately recognizing motions in videos, which can be a valuable tool in applications such as sports analytics and medical diagnosis.

The majority of research in the action recognition field has largely focused on designing deeper and more complex deep neural network architectures for improved accuracy [2, 3]. However, the increasing complexity of deep neural networks has become one of the biggest obstacles to the widespread deployment of deep neural networks on edge devices such as mobile and other consumer devices, where computational [4], memory, and power capacity is significantly lower than that in high performance computing systems.

Convolutional Networks (ConvNets) [5] have recently proven to be quite effective at classifying human actions in images. ConvNets have also been used to handle video-based action recognition problems [6-8]. Deep ConvNets have a lot of modeling capability and can learn discriminative representations from raw visual input using large-scale supervised datasets.

This paper proposes a novel approach to extract spatiotemporal features and implements a Segmental Action Network, a This novel architecture utilizing an I3D model to train on the extracted features. Section 2 covers related works in human action recognition, Section 3 provides detail on the deep model used, Section 4 discusses experimental results, and Section 5 concludes the paper and identifies further work.

2. RELATED WORK

Many approaches have been developed in recent years, but the majority of them have computational limitations, there has been a lot of research based on deep learning to recognize human actions in videos, since videos are 3D spatio-temporal signals, the main idea behind the majority of these studies [9, 10] is to extend Convolutional Neural Networks (CNNs) to include the temporal information contained in videos. Karpathy et al. [11]. proposed several fusion techniques that slightly modify the CNN architectures to operate on stacked video frame inputs. As their results were similar to the results obtained by using individual RGB frames, these techniques were shown to not correctly model the temporal information. In order to operate in the spatio-temporal domain, Ji et al. [12]. proposed a 3D CNN model that performs 3D convolutions on stacked video frames to learn spatio-temporal information between consecutive frames. In addition to the fact that 3D CNNs perform similarly to 2D CNNs, they are computationally expensive to train because they contain many more parameters and do not model long range temporal information.

In the same context, Simonyan and Zisserman [13] proposed a two-stream CNN architecture that learns spatial appearance information from RGB frames and motion information between frames using optical flow [14]. To improve this architecture that considers only a single frame as input, Wang et al. [15, 16] proposed architectures that aggregate the convolutional features at different temporal and spatial positions. However, the streams in these two-stream CNN architectures are independent and there is no shared information between them. These architectures capture only the motion information in short time windows and do not guarantee to keep the most representative features with pooling techniques.

3. FEATURES SELECTION

3.1 Optical flow

Optical flow [14] is the apparent 2D image motion of pixels (Figure 1). The main initial assumption of optical flow is 'brightness constancy assumption', where intensity of pixels in small variations in 'x', 'y' and 't' directions are the same as the original pixel. The brightness constancy equation is given by Eq. (1): The brightness constancy equation.

$$\boldsymbol{I\ x, y, t\ = I(x + dx, y + dy, t + dt)}$$

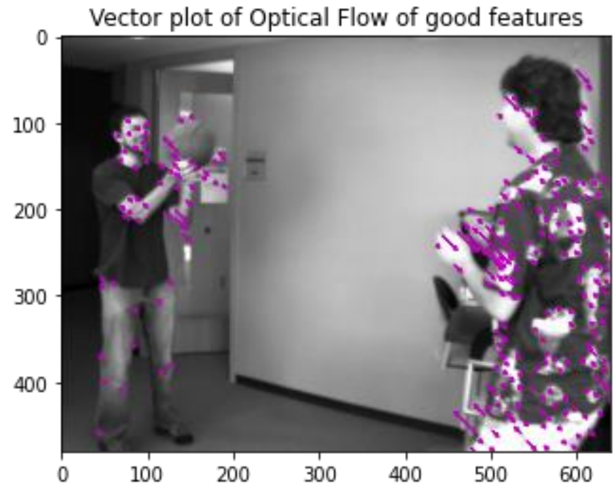

Figure 1. Optical Flow estimation using Lucas–Kanade method

3.2 Space-time interest points

Interest points provide compact and abstract representations of patterns in an image. So, to extend the notion of spatial interest points into the spatiotemporal domain and show how the resulting features often reflect interesting events that can be used for a compact representation of video data as well as for its interpretation.

Many works have been presented to capture STIP information to improve activity recognition.

3.2.1 STIP Methods

The Harris STIP method can be used to detect corners in frames. It is used to detect corners in every pixel of an image by taking into account the corner's differential with respect to direction. If the pixel is in a region of uniform intensity, the adjacent edges will appear comparable. Furthermore, Gabor wavelets are used to find the corners from the exact location of the object using the Gabor STIP.

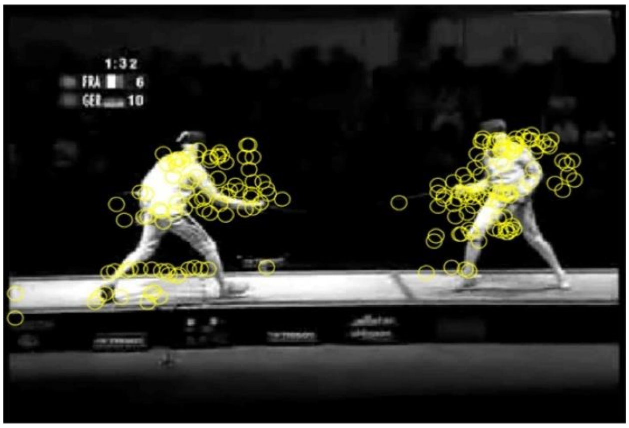

Figure 2. Spatio-temporal interest point detection of two guys fencing

The major goal of the STIPs (Figure 2), according to the study [17, 18], is to perform event detection directly from the image's spatiotemporal data, taking into account regions that have unique locations in space-time with enough robustness to detect and classify.

The idea of the Harris and Forstner interest point operators is used to detect spatiotemporal events by detecting local structures in space-time where the image values have significant local variations in both space and time. Spatiotemporal extents of the detected events are estimated and scale-invariant spatiotemporal descriptors are computed. Video representation is constructed in terms of labeled space-time points by using the descriptors to classify events.

4. PROPOSED APPROACH

The proposed approach divides each video into N segments, with each segment having α frames. The hyperparameter of α is kept fixed for the whole dataset and these chunks are referred to as segments. This approach is split into two stages, with each stage implementing a different neural network.

4.1 Action Frame Selection Network (AFSN)

The main strategy of this architecture (Figure 3) is to extract spatiotemporal features from the entire segment into a single frame, referred to as the "Action Frame".

MobileNet V3 [19], a 2D Convolutional Neural Network (CNN) architecture, has been implemented as the main pipeline for this architecture. It was trained on the ImageNet Dataset, which consists of 14 million images in 1,000 categories. This architecture has strong performance for image classification, object detection, and semantic segmentation tasks, and its characteristics include efficient network architecture, improved network scalability, and improved accuracy on ImageNet.

A pre-trained 2D convolutional neural network is used to extract spatial features from each segment of the video. This network outputs a feature vector for each segment. The top 10 classes of each segment are then predicted and weighted according to their order. A max voting system, referred to as Score Vote, is used to select the frame with the highest score from each segment. This results in the extraction of one frame for each segment that has the highest Score Vote among the other frames in that segment. This process is repeated for each segment of the video. Temporal features are obtained using optical flow, with the previous frame used as the original point for the output action frame.

The Lucas-Kanade (LK) method [20] is a well-known iterative sparse optical flow estimation technique used in computer vision to estimate the motion of objects from successive frames in a video sequence. It is based on the small motion assumption, which enables it to track a set of image points over multiple frames in order to calculate the optical flow. The LK method is highly regarded for its robustness to noise yet requires minimal computations, making it an ideal choice for our application. The Lucas-Kanade method has been selected due to its superior results.

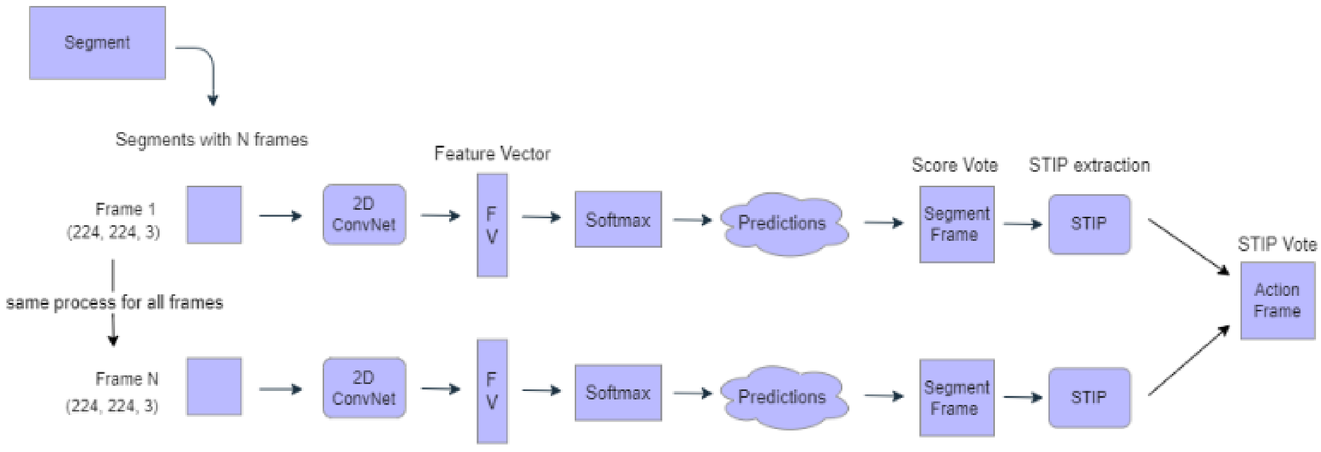

Figure 3. Action Frame Selection Network (AFSN)

4.2 Phase II - Segmental Action Network (SAN)

In the field of image recognition, convolutional neural networks (CNNs) have proven to be an effective feature extractor. It has been demonstrated that just by making the proper adjustments during training, CNNs can achieve much greater success in visual target recognition and classification. Furthermore, CNN has invariance for lighting, and disorderly environmental change. 3D convolutional neural networks were initially developed for an action recognition problem, 3D CNN has been a common research method, as some works [21] shows that 3D CNN is better for low-level spatial temporal features extraction, in this phase, a novel architecture, Segmental Action Network (SAN), is utilized to train the model. SAN implements the Frame Selection Network architecture on each segment to output the selected frame in spatiotemporal dimensions (Figure 4). Subsequently, all frames are concatenated and fed into a pre-trained I3D, a widely adopted 3D video classification network that directly extracts spatiotemporal data from videos using 3D convolution. Consequently, I3D is selected as the main pipeline of the CNNs.

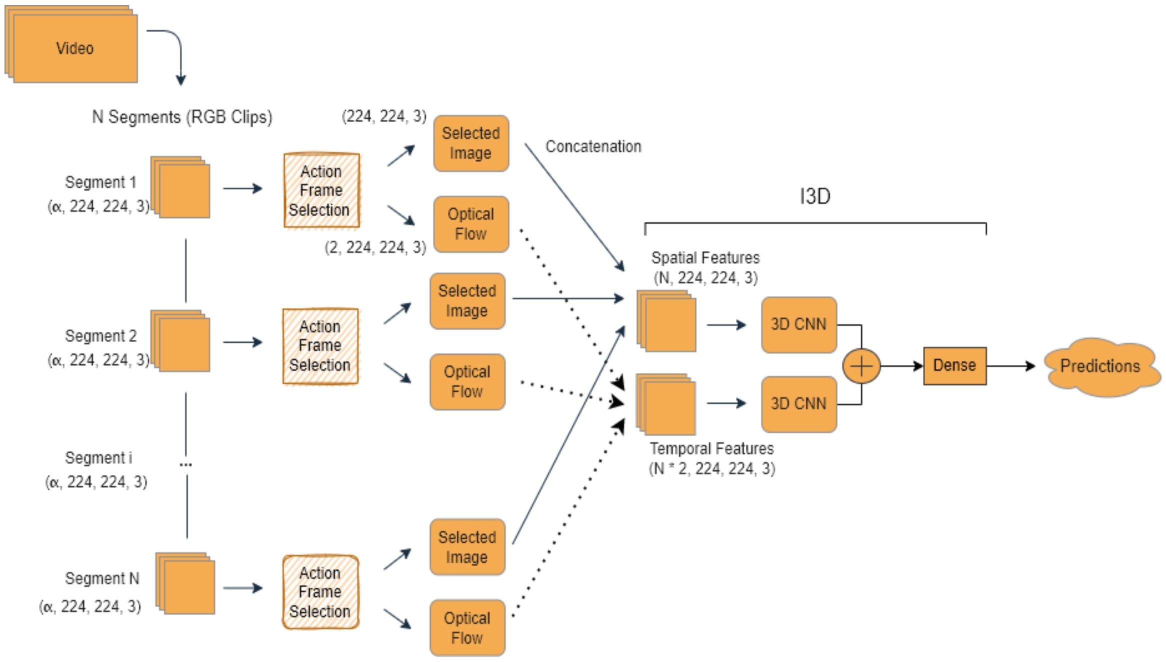

Figure 4. Segmental Action Network (SAN)

5. EXPERIMENTAL RESULTS

5.1 Dataset

The experiments were performed on HMDB-51 [22]. There are 6849 videos in 51 classes in the HMDB-51, with at least 101 clips in each category. The work of classifying each video or action is difficult because all of the videos were acquired from YouTube and contain a wide range of degrees of freedom. These datasets are divided into three groups using various combinations of training and testing data. The original dataset was utilized to train and evaluate the method, ensuring an accurate comparison with earlier methods (Figure 5).

Figure 5. Example of HMDB Dataset action classes

5.2 Experimental settings

Several experiments are conducted to evaluate the performance of our proposed method on HMDB Dataset. The proposed approach is implemented using Python, OpenCV and deep learning API Keras with Tensorflow as a backend. The experiment was performed with an Nvidia Tesla K80 GPU having 4992 Nvidia Cuda cores.

5.3 Results

Multiple pipeline variants for the convolutional networks feature extractor have been implemented, with MobileNet V3 performing well in terms of accuracy among the other backbones.

Table 1. Results of our approach using different 2D CNN pipeline variants on HMDB-51

2D CNN pipeline variant	Accuracy
GoogleNet (Inception) [23]	68.14%
ResNet-50 [24]	70.42%
VGG-16 [25]	71.21%
Xception [26]	72.76%
SqueezeNet [27]	73.18%
MobileNet V3	**73.32%**

A comparison of our model and other models is shown in Table 1. Note that Segmental Architecture Network (SAN) achieved comparable performance by utilizing a smaller backbone network such as MobileNet V3.

The reason for the significant difference in GFLOPs between MobileNet V3 and other popular models in Table 2 is that MobileNet V3 is a very lightweight architecture, whereas the other models have a very large number of parameters.

A comparison of the 3D CNN pipeline is shown in Table 3. I3D was used as the main 3D CNN pipeline due to its performance. the best in all of the studies mentioned, which were all conducted on the same platform.

Table 2. Comparison of our approach using different backbones on HMDB-51, the complexity is evaluated using FLOPs, i.e. floating-point operations per second, results are only using RGB information, no optical flow is used for these experiments

2D CNN pipeline variant	FLOPs
ResNet-50	217 G
VGG-16	124 G
Xception	168 G
MobileNet V3	**74 G**

Table 3. Results of our approach using different 3D CNN pipeline variants on HMDB-51

3D CNN pipeline variant	Accuracy
Two-Stream	52.6%
C3D	64.43%
Res3D	69.18%
TSN	72.18%
I3D	73.32%
MobileNet V3	**73.32%**

5.4 Comparison

The value of the representation flow was verified by comparing it to other CNN-based motion representation techniques currently in use. When the authors' code was accessible, it was utilized for the tests, otherwise the methods were carried out independently. The use of 3x224x224 as input frames enabled more accurate comparisons to previous works. Table 4 show a comparison between our model and other models on HMDB-51.

Table 4. Comparison of our approach with other results on HMDB-51

Method	Backbone	Pre-train Data	Accuracy
C3D	3D VGG-11	Sports-1M	51.6%
STC [28]	ResNet101	Kinetics	66.8%
ARTNet with TSN [29]	3D ResNet-18	Kinetics	70.9%
ECO [30]	BNInception+ 3D ResNet-18	Kinetics	72.4%
TSN [31]	ResNet-50	ImageNet+Kinetics	54.7%
TSM [32]	ResNet-50	ImageNet	70.7%
STM [33]	ResNet-50	ImageNet+Kinetics	72.2%
Our Approach (SAN)	**MobileNet V3**	**ImageNet**	**73.32%**

Table 4 provides a summary of the performance comparisons. The Segmental Action Network scores overall of 73.32% on HMDB-51 show that the visual representation created by our feature extraction method provides an improvement in performance over baselines on the action recognition task.

6. CONCLUSION

This study presented a novel approach to learn better representations for action recognition in videos. The proposed approach extracts the action frame spatial feature from the segments using MobileNet V3, then applies an optical flow to this action frame, which helps to extract the temporal features. This leads to improved performance on the downstream task of action recognition as well as enhancing performance by reducing computational costs, as the video is segmented into chunks to make the system benefit from parallel computing. Furthermore, the proposed architecture was able to effectively extract visual characteristics from each image, calculate the optical flow in the sequence of selected images, and use a 3D spatio-temporal convolutional network as a feature extractor for human action recognition. Results from the HMDB database showed that the proposed model was able to accurately recognize human actions with a high level of accuracy. The approach can be extended to explicitly handle missing joint information and people in the background. there is considerable potential for applications of these methods in a variety of scenarios, and further research is needed to explore these applications.

DATA AVAILABILITY

The datasets analyzed during the current study are available at https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/#dataset.

REFERENCES

[1] Ramanathan, M., Yau, W., Teoh, E. (2014). Human action recognition with video data: research and evaluation challenges. IEEE Transactions on Human-Machine Systems, 44(5): 650-663. https://doi.org/10.1109/THMS.2014.2325871

[2] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S. Vinyals, O., Monga, R., Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. Computer Vision and Pattern Recognitio. https://doi.org/10.48550/arXiv.1503.08909

[3] Sanou, I., Conte, D., Cardot, H. (2019). An extensible deep architecture for action recognition problem. Computer Science, pp. 191-198. https://doi.org/10.5220/0007253301910199

[4] Russakovsky, O., Deng, J., Su, H., Krause, J. Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115: 211-252.

[5] O'Shea, K., Nash, R. (2015). An introduction to convolutional neural networks. Neural and Evolutionary Computing. https://doi.org/10.48550/arXiv.1511.08458

[6] Gan, C., Wang, N., Yang, Y., Yeung, D.Y., Hauptmann, A.G. (2015). Devnet: A deep event network for multimedia event detection and evidence recounting. Computer Vision and Pattern Recognition, pp. 2568-2577. https://doi.org/10.48550/arXiv.1511.08458

[7] Gupta, R.K., Chia, A., Rajan, D. (2013). Human activities recognition using depth images. In Proceedings of the 21st ACM International Conference on Multimedia, pp. 283-292. http://dx.doi.org/10.1145/2502081.2502099

[8] Deng, J., Dong, W., Socher, R., Li, L.J., Li K., Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, FL, USA, pp. 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[9] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, pp. 4489-4497.

[10] Liu, J., Musialski, P., Wonka, P., Ye, J.P. (2012). Tensor completion for estimating missing values in visual data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35: 208-220. https://doi.org/10.1109/TPAMI.2012.39

[11] Karpathy, A. Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), OH, USA, pp. 1725-1732.

[12] Xu, W., Miao, Z., Yu, J., Ji, Q. (2019). Action recognition and localization with spatial and temporal contexts. Neurocomputing, 333: 315-363. https://doi.org/10.1016/j.neucom.2019.01.008

[13] Simonyan, K., Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems 27 (NIPS 2014), pp: 1-8. https://doi.org/10.48550/arXiv.1406.2199

[14] Beauchemin, S.S., Barron, J. L. (1995). The computation of optical flow. ACM Computing Surveys (CSUR), 27(3): 433-466. http://dx.doi.org/10.1145/212094.212141

[15] Yu, Z.X.J., Zhou, F., Qin, J., Ni, D., Chen, S., Lei, B., Wang, T. (2018). Melanoma recognition in dermoscopy images via aggregated deep convolutional features. IEEE Transactions on Biomedical Engineering, pp. 1006-1016. https://doi.org/10.1109/TBME.2018.2866166

[16] Wang, L., Xiong, Y., Wang, Z., Qiao,Y., Lin, D., Tang, X., Gool, L.V. (2016). Temporal segment networks: Towards good practices for deep action recognition. European Conference on Computer Vision, pp. 20-36. https://doi.org/10.48550/arXiv.1608.00859

[17] Yashwanth, K.R., Sunay, M.N., Srinivas, S., Abhishek, Rao, Usha, M.S. (2020). STIP based activity recognition. International Journal of Engineering Research & Technology (IJERT), 8(11). https://doi.org/10.17577/IJERTCONV8IS11050

[18] Harris, C.G., Stephens, M. (1988). A combined corner and edge detector. In Alvey Vision Conference, pp. 10-5244. https://doi.org/10.5244/C.2.23

[19] Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), pp. 1314-1324. https://doi.org/10.1109/ICCV.2019.00140

[20] Baker, S., Matthews, I. (2004). Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision, 56: 221-255.

[21] Ji, S., Xu, W., Yang, M., Yu, K. (2012). 3D Convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35: 221-231. https://doi.org/10.1109/TPAMI.2012.59

[22] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T. (2011). HMDB: A large video database for human motion recognition. In 2011 International Conference on Computer Vision, Barcelona, Spain, pp. 2556-2563.

[23] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, MA, USA, pp. 1-9.

[24] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, NV, USA, pp. 770-778.

[25] Sengupta, A., Ye, Y.T., Wang, R., Liu, C., Roy, K. (2019). Going deeper in spiking neural networks: VGG and residual architectures. Frontiers in Neuroscience, 13: 95. https://doi.org/10.3389/fnins.2019.00095

[26] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, 2017, pp. 1251-1258.

[27] Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K. (2016). Size, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model. arXiv preprint arXiv:1602.07360. https://doi.org/10.48550/arXiv.1602.07360

[28] Diba, A., Fayyaz, M., Sharma, V., Arzani, M.M., Yousefzadeh, R., Gall, J., Gool, L.V. (2019). Spatio-temporal channel correlation networks. Computer Vision (ECCV). https://arxiv.org/abs/1806.07754

[29] Wang, L., Li, W., Li, W., Gool, L.V. (2017). Appearance-and-relation networks for video classification. Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1711.09125

[30] Zolfaghari, M., Singh, K., Brox, T. (2018). Eco: Efficient convolutional network for online video understanding. Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1804.09066

[31] Wang, L.M., Xiong, Y.J., Wang, Z., Qiao, Y., Lin, D.H., Tang, X.O., Gool, L.V. (2019). Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(11): 2740-2755. https://doi.org/10.1109/TPAMI.2018.2868668

[32] Lin, J., Gan, C., Han, S. (2018). TSM: Temporal Shift Module for Efficient Video Understanding. Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1811.08383

[33] Jiang, B., Wang, M., Gan, W., Wu, W., Yan, J. (2019). STM: SpatioTemporal and Motion Encoding for Action Recognition. Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1908.02486