Human Action Recognition in Video Using DB-LSTM and ResNet

Akram Mihanpour , Mohammad Javad Rashti*, Seyed Enayatallah Alavi
Department of Computer Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran
a-mihanpoor@stu.scu.ac.ir, {mohammad.rashti, se.alavi}@scu.ac.ir

***Abstract*— Human action recognition in video is one of the most widely applied topics in the field of image and video processing, with many applications in surveillance (security, sports, etc.), activity detection, video-content-based monitoring, man-machine interaction, and health/disability care. Action recognition is a complex process that faces several challenges such as occlusion, camera movement, viewpoint move, background clutter, and brightness variation. In this study, we propose a novel human action recognition method using convolutional neural networks (CNN) and deep bidirectional LSTM (DB-LSTM) networks, using only raw video frames. First, deep features are extracted from video frames using a pre-trained CNN architecture called ResNet152. The sequential information of the frames is then learned using the DB-LSTM network, where multiple layers are stacked together in both forward and backward passes of DB-LSTM, to increase depth. The evaluation results of the proposed method using PyTorch, compared to the state-of-the-art methods, show a considerable increase in the efficiency of action recognition on the UCF 101 dataset, reaching 95% recognition accuracy. The choice of the CNN architecture, proper tuning of input parameters, and techniques such as data augmentation contribute to the accuracy boost in this study.**

Keywords*— *Action Recognition; Video Processing; Deep Neural Networks; Convolutional Neural Network; DB-LSTM.

I. INTRODUCTION

Today, videos have become the primary and most influential means of communication and socialization. Videos are very popular in prevalent social networking platforms, many of them containing human actors. Surveillance and security cameras are everywhere, and news feeds are hardly credible without an accompanying video. Much valuable information can be extracted from such videos, which may raise our social and national security levels, save and enhance people's lives, and increase productivity in our jobs. Obviously, manual extraction and analysis of such mines of big data is out of question when even automated intelligent systems are struggling to keep pace with their production velocity. Automatic detection and classification of human actions is a popular approach in turning videos to valuable and timely information.

Action recognition can be defined as the ability to automatically recognize a specific activity in a video stream. With its huge application spectrum, human action recognition (HAR) is proven to be an important part of computer vision research[1][2]. These applications include security, surveillance, smart objects, video retrieval, video-content-based monitoring, human motion synthesis for man-machine interaction, sports analysis, health/disability care, and entertainment [3].With challenges such as similarity of visual content, view point and camera movements (w.r.t. the actor), scaling, gesture variation, and variable imaging and lighting conditions, HAR is certainly facing a very challenging research path [4].

State-of-the-art HAR algorithms[5][6] remarkably reduce the human labor in analyzing large-scale video data [3]. With the emergence of deep learning approaches in AI systems, we can extract even more complicated features and concepts of a video stream with higher accuracy and in shorter time. Recently, Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) networks have shown great success in image/video classification and voice recognition [7][8][9]. In this study, we propose a new method for HAR by processing only raw video frames using a pre-trained CNN called ResNet152 and Deep Bidirectional LSTM Network (DB-LSTM). The implementation results using the UCF101 dataset show that the proposed method has a very high performance in HAR in video.

This paper is organized as follows: Section II provides an overview of previous related work. Section III presents the proposed deep learning architecture, with the implementation details laid out in Section IV. In Section V, we will report and analyze the experimental results and finally conclude the paper in Section VI.

II. RELATED WORK

In this section, we classify and overview state of the art in video-based human action recognition in various subsections.

A. Classic Models

Classic models are based on handcrafted features and typically use local ones. There are generally two main classical approaches [10]. A number of classic methods use holistic representations[11], as a global representation of the human body and its movements while some methods use local features. Space-Time Volumes (STVs), which are among the global representation methods, were first introduced in [12].

Many studies have investigated the problem of action recognition using spatio-temporal features[13][14]. One of the basic spatio-temporal models is presented in [15], which provides a method for extracting key spatio-temporal points, i.e., regions of the image that are prominent and sudden changing regions in both space and time dimension.

A better approach is to use dense optical flows and classify these flows into motion bubbles [16], which, of course, tends to be very costly and time-consuming, nevertheless yielding significantly higher accuracy [17][18]. The authors in [19] present the use of a dense sampling method to create a dense trajectory. This technique fuses several different features with a particular structure and is one of the strongest methods among the classic models. Despite being very slow, this method provides a high

978-1-7281-1051-6/20/$31.00 ©2020 IEEE

accuracy level and early deep-learning HAR solutions have been compared to it as a base state-of-the-art method in classical HAR research.

Despite the relatively acceptable performance of handcrafted-features-based approaches, research has recently been more interested towards the use of deep learning models, due to many difficulties of classical approaches, including their low flexibility in intra-group variations of actions.

B. Deep Learning Models

Deep learning models can extract more complicated concepts from videos by considering frame sequences. Many efforts have been made to involve the time information available in the video into CNN models, including the work in [20]. Deep-learning action recognition methods can be grouped in two categories: two-stream networks and space-time networks. Most existing deep-learning HAR methods use two-stream networks[21], one stream for spatial information and the other for inter-frame motion information. In [21], the frames are fed into a two-stream network with two inputs. The raw frame images are fed into the spatial stream and the optical flows images are fed into the temporal stream.

Three dimensional CNNs are examples of space-time networks introduced by Jay et al. [22]. Other space-time networks such as Recurrent Neural Networks (RNN), such as LSTM, have also been used to incorporate temporal information into the video [23][24].

One of the drawbacks of two-stream approaches is that the motion information is processed separately from the visual information. Compared to the two-stream approaches, both 3D and recursive networks demonstrate high processing volumes and have a high number of training parameters, hence the need for very large datasets to train such networks. Since the production of such video datasets is onerous and costly, there is a need for methods that can optimally incorporate temporal information into the processing without requiring large training datasets. The method presented in this paper incorporates temporal information appropriately into processing while requiring less training data than spatio-temporal networks.

III. The Proposed HAR Algorithm

A machine learning process design can be divided into four main parts: the data, the model, the cost function and the optimization function. In the following, we will design the proposed HAR method using these components.

A. Data

The UCF101 dataset is one of the most popular action recognition datasets, with real-world videos. This dataset contains 13320 videos taken from YouTube divided into 101 classes, including music playing, makeup, sports (football), brushing, among others. The UCF101 dataset contains a wide range of actions from the five main categories: human-object interaction, body movement, human-to-human interaction, playing musical instruments, and sports. Each category contains 100 to 200 videos. The shortest video contains 28 frames and the dimensions of each frame is 320*240 [25]. Some categories such as sports are further divided into sub-classes. Most sports videos have a similar (green) background, which makes it challenging to distinguish among the sports. Videos also have varying lighting conditions, actor gestures and viewing points. This is in addition to the sub-optimal quality of many videos in the dataset.

We first need to pre-process the videos by extracting the input video frames before inputting them to the proposed model. Libraries like OpenCV or FFmpeg can be used to extract video frames. We directly imported the pre-processed dataset used in [26]. In the UCF-101 dataset, action tags are non-numerical, thus requiring an encoding step. Label encoding is used due to its much lower number of zeros compared to the one-hot encoding, thereby reducing the overall network training overhead.

B. The Proposed Model

Structure of the proposed model is illustrated in Fig 1. The proposed model for the present study consists of two parts. In the first part, we use a CNN network for categorizing images and videos. The desired features are extracted from each video frame by the CNN network. The output of the CNN network is a feature vector whose dimensions are reduced after the undesirable features are removed by the CNN.

Training a deep learning model to display images requires thousands of images and a high processing power to adjust the weights of the CNN model. To avoid extremely long training times, we use learning transfer where a pre-trained model is employed. We use a pre-trained CNN network named ResNet152 [27]. ResNet is pre-trained on

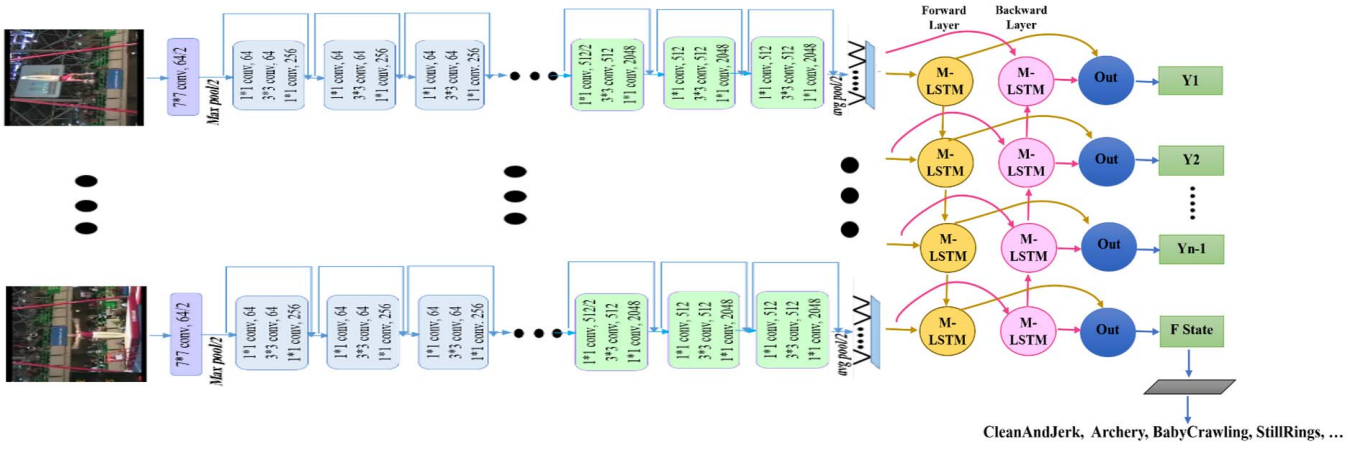

Fig. 1. Structure of proposed method for action recognition

a very large ImageNet[28] dataset with over 15 million images.

ResNet152 consists of 152 layers, including 150 convolutional layers, a pooling layer and a fully connected layer. In very deep architectures such as ResNet and its variants, the number of learnable parameters increases with depth. This increases the computational overhead, and, as a result, reduces the speed of network training and learning. To avoid this problem, a technique called bottleneck is used in ResNet architecture. After the first two layers, the input space is compressed and the input with 224*224 size is reduced to 56*56. In each block, the input is then processed by three layers of convolution. ReLU and batch normalization are used in each layer. After normalizing the batches in each layer, the ReLU nonlinear activation function better updates the weights in each layer, resulting in faster learning time and lower computational overhead. Batch normalization is also a technique for improving the performance and stability of neural networks. The idea is to normalize the input of each layer with mean and variance. By normalizing data, learning and reducing error rate are faster.

In frame sequences, a frame does not only depend on its previous frames, but may also depend on its subsequent ones. In this case, to increase the accuracy and efficiency of the model, it is advisable to have a structure that processes the input frames in both directions. For this purpose a Deep Bidirectional LSTM (DB-LSTM) network is employed in the proposed model. According to the DB-LSTM structure, two LSTM networks simultaneously process input from both directions, one reads and processes the input forward, and the other reads and processes the input backward. DB-LSTM helps detecting complex and sequential patterns hidden in video frames.

The output of each layer in the bi-directional RNN represents the action performed in the frame processed in that layer. Finally, the input video tag is determined by the maximum frequency of the frames tag. In the classification step, the highest score tag is displayed as the output and the action recognized in the input video.

C. Cost Function

The proposed model is expected to learn properly in order to correctly identify and differentiate various actions in videos. To evaluate the model, we use a cost function, aiming at measuring the model performance and maximizing it by minimizing the cost.

The cost quantity is defined as the amount of actual output difference from the network's expected output. There are various cost functions for classification problems, one of the most famous ones (that usually works well in classification problems) being the Cross Entropy function [29] as defined in equation (1):

$$\mathrm{CE}(y,s) = -\sum_c y_c . \log(s_\theta(x)_c) \qquad (1)$$

In (1), c represents the class, yc is the actual value of the class and $s_\theta(x)_c$ is the output value of the model designed for input x.

D. Optimizer Function

To minimize the cost function an optimization algorithm is required. There are various optimization algorithms, with which the proposed model was tested, including SGD, StepLR and Adam. Based on our experimental results, the Adam optimizer function yields a better performance and accuracy compared to other optimizer functions.

IV. Implementation

All parameters of the proposed model are obtained by performing various tests and evaluating the accuracy and speed of the proposed model training with different values of the parameters. We chose 1×10^{-3} for the learning rate, then tested the model with different batch sizes. Considering the accuracy and the available memory on the platform, we chose 128 for the batch size parameter.

For data augmentation [30], we performed processing on the network input frames using three strategies, namely RandomRotation (5), transforms.RandomCrop (crop _ size) and Random Horizontal Flip, and the latter yields the highest accuracy. With these three transformations, from each frame we produce three additional frames with a slightly different look but the same label. This technique has enabled our proposed model to learn robust and differentiated data, leading to higher accuracy.

Since transferring data from disk to GPU is very costly and time consuming, we use multi-threading in order to make the model execution faster. Particularly, four threads are used. Using this technique increases the efficiency compared to the single thread case.

V. Experimental Results

For testing and training, the dataset is divided into two parts: 75% for training and 25% for testing. Part of the training data are randomly and variably used for validation. The reported final evaluation is the accuracy obtained in action recognition on the test data. This accuracy is the percentage of matching between the actual value of the video tag in question and the value predicted by the proposed model.

The library used for image and video processing is Torch. Furthermore, we used PyTorch framework for machine learning and deep learning functions. The simulation in this research has been done in 120 iterations on a machine with a 1080 NVIDIA GeForce GTX GPU, and 8 GB of GPU RAM.

For a fair evaluation of the proposed method, the results are compared to the state-of-the-art research, as outlined in Table I. The results table shows the average recognition score obtained in this study compared to other similar studies. As the results indicate, the accuracy of the proposed method is higher than the previous state of the art, particularly when compared to the two-stream methods [21], with one flow on raw video frames and the other on optical flow frames. The accuracy is also higher than the CNN based methods such as deeply-transferred motion vector [31], 3D CNNs [32], and factorized spatio-temporal CNNs[33]. Moreover, it is more accurate and faster than the study in [34], despite its smaller number of iterations.

Fig 2 presents the accuracy and loss of the training and test steps obtained at each iteration of the implementation. As is observed in this figure, when the number of implementation steps increases, the accuracy increases and the loss decreases, an evidence that the training is well done and no overfiting is present.

The achieved accuracy and speed can be attributed to the use of ResNet152 network, multi-threading, and data

augmentation. The method uses a pre-trained ResNet152 network to extract frame features, followed by a bidirectional recursive network with long-term memory to extract sequence information between frames. Using such a network enables the sequences of video input frames to be processed in both directions (forward and backward), thereby increasing the probability of an accurate action classification. By the use of techniques such as data augmentation, this is further improved, while reducing the probability of overfiting. Moreover, the proper selection of input parameter values is another contributor to the accuracy score, when compared to other similar methods.

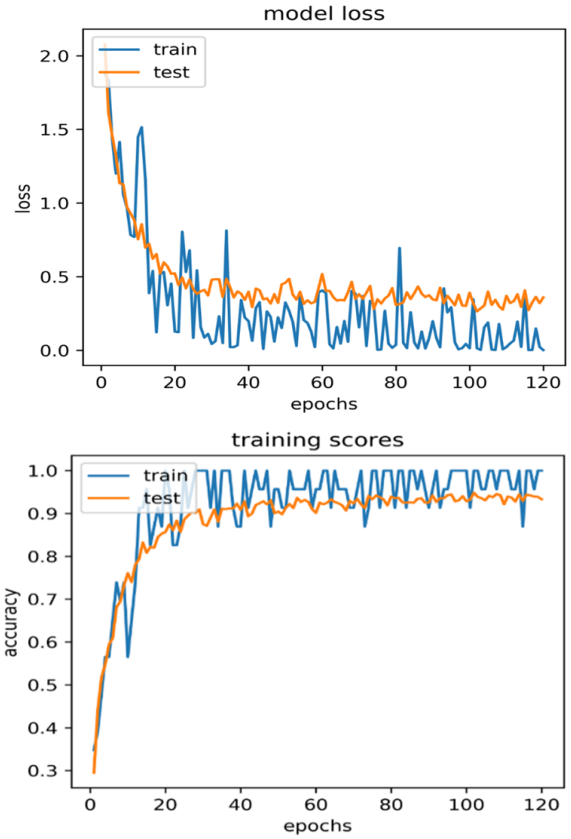

Fig. 2. Accuracy and loss diagrams of the proposed model in both training and testing steps

The confusion matrix for 40 classes of the split1 of UCF101 dataset is presented in Fig 3. For example, in this matrix, the class "HairCut" in our method achieves 97% true positive prediction and the class "BaseballPitch" has only 1% false prediction. As observed, the intensity of true positives (diagonal) is high for each category, indicating the high capability of the proposed method for human action recognition in the video. Overall, we observe an average of about 95% accuracy (see Table I) among the classes across all splits of the dataset.

As for processing speed, with multi-threading, the system takes approximately 0.11 sec for feature extraction per frame. Feeding the extracted features to DB-LSTM for classification takes 0.47 sec for 30 frames per second video clip. Overall, our method takes approximately 0.58 seconds for processing of a 1-second video clip. With these statistics, our method can process 52 frames per second, making it a suitable candidate for action recognition in real-time video processing applications.

VI. Conclusion

In this study, a new method is presented for human action recognition using CNN and Deep Bidirectional LSTM (DB-LSTM) using only raw video frames (instead of using optical flow data). First, deep features of the video frames

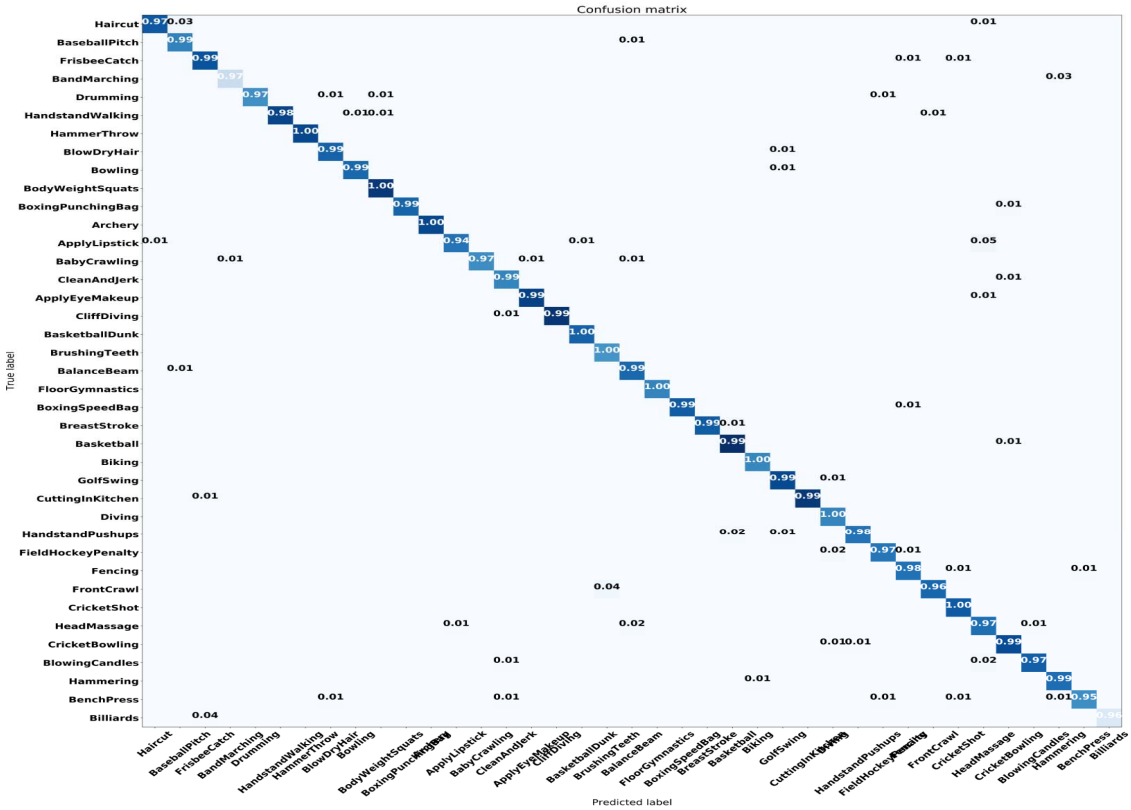

Fig. 3. Confusion matrix of UCF101 dataset for the proposed HAR method

TABLE I. COMPARISON OF AVERAGE RECOGNITION SCORE OF THE PROPOSED METHOD WITH PREVIOUS METHODS

Method	UCF-101 Accuracy
C3D(3net)[32]	85.2%
Two-stream CNNs[21]	88.0%
EMV+RGB-CNN [35]]	86.4%
RLSTM-g3[36]]	86.9%
Multiple dynamic images [37]]	89.1%
Factorized spatio-temporal CNNs [33]]	88.1%
Temporal pyramid CNNs[33]]	89.1%
DTMV+RGB-CNN[31]]	87.5%
DBLSTM+CNN[34]]	91.21%
Our method (ResCNN-DBLSTM)	94.79%

are extracted using the ResNet152 pre-trained convolutional Neural Network architecture, to accelerate the learning process and improve the performance. Then, the sequence information of the frames is learned using the DB-LSTM recurrent network in both forward and backward transitions, and the final classification is performed.

The simulation results using PyTorch on a GPU-equipped machine show that the proposed method has a very high performance in HAR from video on the UCF101 dataset, compared to state-f-the-art methods. Proper adjustment of input parameters, the use of the pre-trained ResNET152 and techniques such as data augmentation have increased the accuracy in this study. Moreover, the use of multiple threads in parallel, for data preparation and injection reduces the training time. This makes our method more appropriate for real-time visual data processing and can be used as an integral part of smart detection systems.

ACKNOWLEDGMENT

This study was funded by Shahid Chamran University of Ahvaz (grant number SCU.EC98.30899). The authors would also like to thank SCU's Deep Learning Laboratory, the Department of Computer Engineering for the computing resources used for this research.

REFERENCES

[1] S. Saif, S. Tehseen, and S. Kausar, "A survey of the techniques for the identification and classification of human actions from visual data," *Sensors,* vol. 18, no. 11, p. 3979, 2018.

[2] S. Ranasinghe, F. Al Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," *International Journal of Distributed Sensor Networks,* vol. 12, no. 8, p. 1550147716665520, 2016.

[3] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv preprint arXiv:1806.11230,* 2018.

[4] R. K. Tripathi, A. S. Jalal, and S. C. Agrawal, "Suspicious human activity recognition: a review," *Artificial Intelligence Review,* vol. 50, no. 2, pp. 283-339, 2018.

[5] Y. Kong, S. Gao, B. Sun, and Y. Fu, "Action prediction from videos via memorizing hard-to-predict samples," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*, 2016: Springer, pp. 20-36.

[7] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 40, no. 6, pp. 1510-1517, 2017.

[8] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo, "3d human activity recognition with reconfigurable convolutional neural networks," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 97-106.

[9] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream lstm: A deep fusion framework for human action recognition," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017: IEEE, pp. 177-186.

[10] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing,* vol. 60, pp. 4-21, 2017.

[11] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* no. 3, pp. 257-267, 2001.

[12] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2005, vol. 2: IEEE, pp. 1395-1402.

[13] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," 2009: Citeseer.

[14] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition*, 2009: IEEE Computer Society, pp. 2929-2936.

[15] I. Laptev, "On space-time interest points," *International journal of computer vision,* vol. 64, no. 2-3, pp. 107-123, 2005.

[16] R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998: IEEE, pp. 416-421.

[17] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European conference on computer vision*, 2004: Springer, pp. 25-36.

[18] V. Kantorov and I. Laptev, "Efficient feature extraction, encoding and classification for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2593-2600.

[19] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551-3558.

[20] B. Fernando and S. Gould, "Learning end-to-end video classification with rank-pooling," in *International Conference on Machine Learning*, 2016, pp. 1187-1196.

[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568-576.

[22] 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.

[23] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110-1118.

[24] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*, 2015, pp. 843-852.

[25] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402,* 2012.

[26] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933-1941.

[27] Y. Bengio, "Deep learning of representations: Looking forward," in *International Conference on Statistical Language and Speech Processing*, 2013: Springer, pp. 1-37.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[29] N. T. Vu, P. Gupta, H. Adel, and H. Schütze, "Bi-directional recurrent neural network with ranking loss for spoken language understanding," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016: IEEE, pp. 6060-6064.

[30] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[31] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with deeply transferred motion vector cnns," *IEEE*

Transactions on Image Processing, vol. 27, no. 5, pp. 2326-2339, 2018.

[32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489-4497.

[33] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4597-4605.

[34] 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.

[35] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with enhanced motion vector CNNs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2718-2726.

[36] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3D human-skeleton sequences for action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3054-3062.

[37] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3034-3042.