Two-Stream 3D Convolution Attentional Network for Action Recognition

Raden Hadapiningsyah Kusumoseniarto
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C.
Email: M10615811@mail.ntust.edu.tw

***Abstract*—We propose a new method, which uses a two-stream 3D convolution network to capture rich spatial and temporal information, then process it with an attention module to capture long- and short-term dependency, to recognize action on the videos. By taking advantages of 3D convolutions, not only spatial information is obtained, but the movement information on the videos is also captured as temporal information. The main reason to consider long-term temporal dependency information is that it will be important to identify action on the videos. The bidirectional self-attention network uses forward/backward masks to encode temporal order information, and attention to handle our sequence on 3D convolution features. We employ a two-stream 3D network to capture spatial-temporal features, combined with self-attention to further capture temporal relationships.The experimental results indicate that the proposed method can be compared to state-of-the-art work in the HMDB-51 dataset with a less complex process while maintaining the performance.**

We employ a two-stream 3D network to capture spatial-temporal features, combined with self-attention to further capture temporal relationships.

***Index Terms*—3D convolution, attention module, action recognition**

I. Introduction

Video understanding is getting a lot of attention due to its broad applications. For instance, action recognition [1], video surveillance [2], face recognition [3], robotics [4], action localization [5], *etc.* However, recognizing the actions in the videos poses numerous challenges, such as the inter-class variation problem, occlusion, and background clutter. It is a difficult task to establish efficient and reliable feature representation in order to effectively identify the behaviors.

There have been algorithms proposed for action recognition [6, 7, 8, 9, 10]. For instance, Feichtenhofer *et al.* [11] proposed a two-stream network with a convolutional fusion layer between networks and a temporal fusion layer to process spatio-temporal information. Tran *et al.* [12] separated 3D convolution into 2D spatial convolution and 1D temporal convolution to balance the trade-off between accuracy and complexity. To better capture motion information, several works have been coorporating with spatial information along with motion information [6, 8, 9]. For example, Crasto *et al.* [7] proposed a motion augmented spatial stream to better capture motion information with more efficient computational costs. Choutas *et al.* [8] utilized pose information to encode the movement information through semantic keypoints. Also, Piergiovanni *et al.* [9] stacked multiple representations of flow layers to learn motion information and any representation channels on each CNN layer. Liu *et al.* [6] used recurrent attention mechanisms to adaptively render more informative joints features for 3D action recognition. Wang *et al.* [13] introduced Hallucinating Improving Dense Trajectory that extracted via Bag of Words and Fisher Vector that integrate with I3D model to capture motion information. However, the methods mentioned above [6, 8, 9] did not reckon long-term dependencies information.

978-1-7281-9331-1/20/$31.00 ©2020 IEEE

In this paper, we present a two-stream network combined with a self-attention module to capture both long- and short-term dependencies among spatio-temporal information for action recognition. The proposed method first extracts the RGB and the optical-flow images from the videos. Thereafter, we employ 3D convolutional I3D network [1] ti generate discriminative spatio-temporal features information. An attention module is then adopted to render an important segment from sequence 3D convolutional features. Finally, a two-stream 3D convolutional attention features is concatenate to be trained to model both long- and short-range dependencies.

The contributions of this paper are as follows:

1) we propose a two-stream 3D network to better capture motion for action recognition;
2) we utilize the self-attention module to capture long- and short-range dependencies on spatio-temporal features;
3) compared with state-of-the-art studies, our method uses a less complex process while maintaining promising result.

II. Related Work

Action recognition on the videos has been investigated based on the handcrafted visual features, such as HOG/HOF [14], HOG3D [15], MBH [16] features extracted along dense trajectories. Sun *et al.* [17] proposed a method that utilized both local descriptor and holistic information to recognize actions in the videos. Sadanand and Corso built ActionBank [18], which consists of numerous individual action detectors widely sampled in semantic space for action recognition. Baumann *et al.* [19] introduced motion binary pattern that extract the motion information to identify the actions. However, these approaches [14, 15, 16, 19] have difficulties in handling noise, brightness problem, and requires capabilities to built feature detectors, descriptors, and vocabulary construction methods.

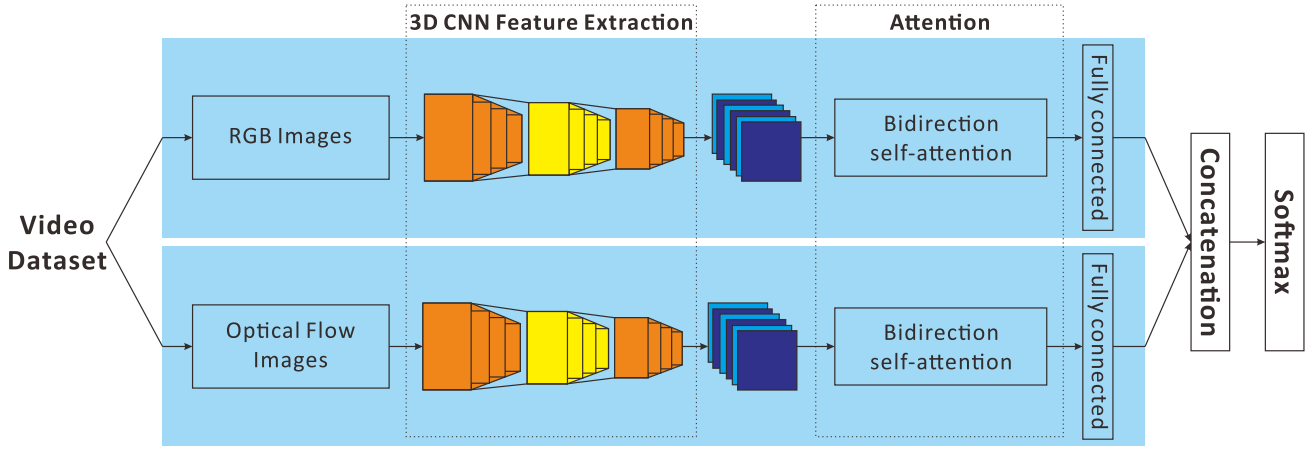


Fig. 1: Architecture model of our proposed method: we train our spatial and temporal 3D convolutional network independently and process to attention module for capture long- and short-term dependency. Then, we concatenate our two-stream network and train with Keras.

Spatio-temporal methods are often implemented for action recognition in terms of detecting spatial and temporal features. Bobick and Davis [20] combined motion history images and motion energy images techniques, which transformed multiple frames to a single image for recording motion in the video. Gorelick *et al.* [21] captured the spatio-temporal features, such as local space-time salience, dynamics of movement, configuration of form, and orientation. Noguchi *et al.* [22] introduced spatio-temporal information combined with multiple kernel learning to recognize actions in the videos. Zhang *et al.* [23] developed slow feature analysis to extract valuable motion pattern information as slowly diversifying features from a rapidly changing action video.

Deep learning based methods automatically generate optimized features from videos. Simonyan and Zisserman [24] discovered that 2D convolution network with deeper architecture gained good performance. 2D convolution network can generate spatial information from single slice as input. However, it fails to leverage background from adjacent or sequence slices. Tran *et al.* [12] applied 3D convolution network for preserving temporal information of 2D convolutional network. Wang *et al.* [25] introduced trajectory-pooled deep-convolutional descriptor that captures a set of point trajectories and inherits the characteristic of deep learning to acquire discriminative convolutional feature maps. Ji *et al.* [26] developed a 3D convolution neural network to obtain spatial and temporal information, then captures the motion information from multiple frames of the video. The 3D convolution network produces a volume feature that captures temporal information due to processing multiple frames at once. 3D convolutional neural network that provides spatial and temporal information can be used for action recognition in the videos.

Shen *et al.* [27] created the bi-directional block self-attention network that works with recurrent neural network or convolutional neural network sequence encoding. Qin *et al.* [28] created a dual-stage attention-based recurrent neural network to adjust the relevant time series input feature in each time step by referring to the previous hidden state. Pramono *et al.* [5] implemented bi-directional block self-attention module for action localization. Wang *et al.* [29] introduced hierarchical attention network to better capture short- and long-term spatial features on the videos for action recognition. Wang *et al.* [30] proposed residual attention networks to capture mixed attention and this is an extensible convolution neural network. Zhang *et al.* [31] introduced self-attention generative adversarial network to make self-attention module integrate into generative adversarial network framework, which is effective in modeling long-range dependency information. With all the sequences of spatial and temporal features of 3D convolutional neural network, our work can be improved with bi-directional block self-attention network to capture long- and short-term dependencies.

III. Proposed Method

In this section, we first introduce the 3D convolution feature extraction as our input of self-attention module. Then, we define our self-attention network to produce our attention features and passed it to fully connected layer. Finally, we concatenate each of our two-stream attention feature and softmax layer to predict action class scores.

A. 3D Convolution Feature Extraction

To be able to learn object movements, a set of spatio-temporal features was generated from intermediate 3D convolutional network [32]. We use the I3D [1] then take advantage of huge pre-trained model such as Kinetics to generate our 3D Convolution features. The I3D model consists of 16 layers, which are 3 convolution layers, 4 max pool layers, and 9 inception module layers. For our spatial and temporal

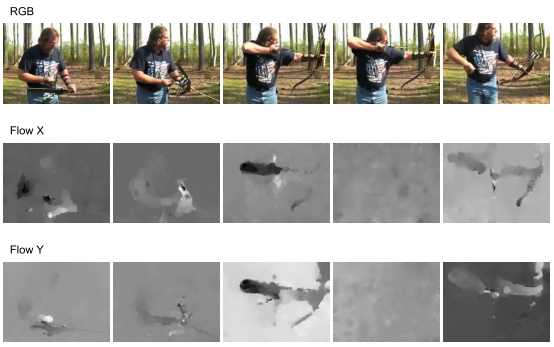


Fig. 2: The visualization of different input modalities, where the top row is RGB, followed by optical-flow (x,y-directions).

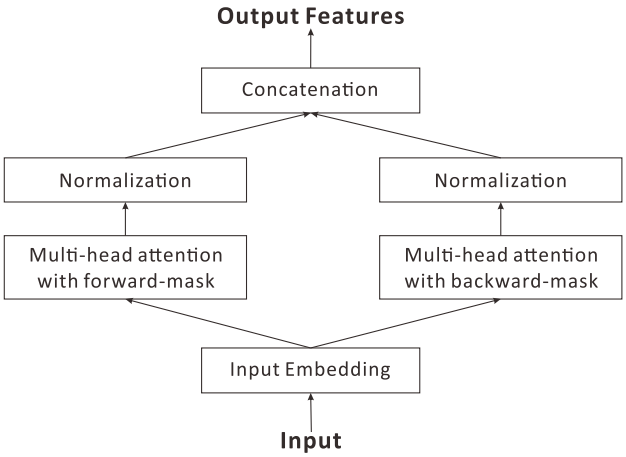


Fig. 3: Bidirectional self-attention network.

networks, RGB and optical-flow were produced as shown in Fig. 2. The optical-flow is generated by TV-L1 algorithm [33].

Traditional 2D convolution neural networks only provide spatial information [32], which lacks the information of the movement information. For video understanding, the movement information is essential to be able to better capture the action from the videos. The creation of spatio-temporal features is a crucial task, since it includes scene-related details and video action, making it useful for different tasks [32]. With 3D convolutional neural networks, convolution and pooling operations are handled spatio-temporally while in 2D convolution neural network, they are performed only spatially from the local neighborhood on feature maps in the previous layer. The temporal stream network takes multiple frames as input, each time using of 2D convolutional layer it slightly eliminates the temporal information [32].

B. Self-attention Network

This is inspired by the successful of attention module that able to learn long-term dependencies in natural language processing area [34]. Self-attention is a method for comparing various points of a single sequence to determine a representation of the same sequence, which is flexible in modeling both long- and short-range dependencies [35]. We can see 3D convolution features as sequence of 2D convolution information that consist of spatio and temporal information.

As shown in 3, bidirectional self-attention network uses forward/backward masks to encode temporal order information, and attention to handle the sequence on 3D convolution features. The purpose of using bidirectional self-attention is to perform an efficient task as a recurrent neural network and acquire all the advantages of self-attention network [27]. The process of bidirectional self-attention network is a concatenation of two multi-head attention modules followed by input sequence of token embedding. One with forward-mask and another with backward-mask, where several soft-attention layers run in parallel.

The attention features are then produced using [27]. First, we extract 3D CNN features of RGB and optical-flow on each videos. Then, we process the features to multi-head attention, one with forward-mask and the other with back-ward mask where each mask of multi-head attention captures long-range/global dependencies. The multi-head attention processes the input through self-attention network multiple times in parallel. To combine the local and global context features, we dynamically merge the input and output of multi-head attention. Afterwards, we obtain the long-range context representation as the output of each masked multi-head attention. Lastly, we concatenate both output of multi-head attention forward-mask and multi-head attention backward-mask as the attention features. Finally, the output of the two-stream is attention features passed it to the fully connected layer. We concatenate it and put it to softmax layer to generates action class scores.

IV. Results and Discussion

A. Dataset and Setup

We evaluate our proposed method on a popular benchmark dataset, HMDB-51, which consists of 6766 video clips with 51 different of activities, and each video has 20-1000 frames [36]. The action labels include *brush hair*, *catch*, *eat*, *fencing*, *hug*, *jump*, *push up*, *ride bike*, and *run* etc. Each video has a single label for both training and testing. The training and testing split sets contain 3570 and 1530 videos respectively.

B. Experimental Setup

Since our action recognition framework composed by 2 networks, we train it one after another, initialized by the training of I3D network for 3D convolution feature extraction and followed by the training of bidirectional self-attention for action recognition. For 3D Convolution network's training, we set the learning rate as 0.0001, batch size as 4, the number of frames for each clip is 64, optimized by Adam optimizer, and the number of steps is 30,000. For the attention scheme we set number of heads attention as 128, and drop rate as 0.9. We concatenate the output of final connected layers and

TABLE I: The recognition performance of I3D with and without attention scheme in HMDB-51 dataset, where the best recognition results are bold-faced.

RGB	Flow	Self-attention	Accuracy
✓	-	-	73.1
-	✓	-	75.2
✓	✓	-	80.0
✓	✓	✓	**81.7**

TABLE II: The performance of the proposed method compared with the state-of-the-art works in HMDB-51 dataset, here the best recognition results are bold-faced. The reproduced result is indicated by †.

Baseline HR	Modalities	Accuracy
Spatiotemporal Convolutions [12]	RGB	78.7
Motion-Augmented RGB Stream [7]	RGB, Flow, MARS	79.5
I3D [1]	RGB, Flow	80.7
PoTion [8]	RGB, Flow, Pose	80.9
Representation Flow [9]	RGB, Flow	81.1
Global Context-Aware Attention [6]	Pose	81.1
Discriminative Pooling [38]	RGB, Flow	81.3
Hallucinating IDT [13]	RGB, Flow	**82.5**
I3D [1] †	RGB, Flow	80.0
Ours	RGB, Flow	81.7

use softmax layer to generates action class scores using Keras library [37] with the learning rate set as 0.000001, regularizer with L2 normalization, optimized by Adam optimizer and number of epochs is 40,000. We assess our proposed method using the accuracy metric is provided by [36].

C. Ablation Studies

Impact of the Combination of Each Module: We also investigate the performance of the combination of each module, as shown in Table I, from which we can note that optical flow is around 2.1% better than RGB input modality. This is because the motion cue information, which is essential for action movement, can be better captured with optical flow modality. The combination of those modalities noticeably improves the performance by about 4.8%, due to its capability to learn spatial and temporal movement on the videos. Finally, combined with attention scheme, we can boost the performance by about 1.6% as this combination can learn the long- and short-range dependencies among the frames.

Confusion Matrix: We visualize our best performance into a confusion matrix, shown in Fig. 5, which explains that our actions are well recognized. The worst recognition is the sword fighting, where only 9 correct predictions out of 30 videos, as this action is similar to fencing and sword exercise. The second worst prediction is hand waving because this action has resemblance with walking, which showed little hand movement and people standing. The third worst prediction is jumping, as this action shows people above the ground and is similar to the somersault.

Fig. 4 shows the best recognized actions and the worst recognized actions. Each left image of Fig. 4a, 4b, 4c, 4d is RGB images, and the right image is feature visualization. Fig. 4a and 4b shows the actions that our model well recognized, and Fig. 4c and 4d shows the actions that our model is unable to fully grasp their essence.

D. Comparison with State-of-the-art Work

We compare our proposed method with eight baseline methods, Spatiotemporal Convolutions [12], Motion-Augmented RGB Stream [7], I3D [1], PoTion [8], Representation Flow [9], Global Context-Aware Attention [6], Discriminative Pooling [38], and Hallucinating IDT [13]. The results of the baselines are obtained from the literature.

From Table II, we can note that [12] attains the worse performance as this method only employed RGB modalities, which cannot capture the motion information. The motion-augmented RGB stream [7] obtains better performance due to the ability to capture motion information. [1] provides better performance by learning spatio-temporal features with applying 3D convolution by feed the network with multiple frames at once. [8] obtains a slight improvement with the additional pose information to capture the local movement. [9] outperforms the aforementioned approaches by employing salient motion features from representation of optical flow. [6] obtains the similar results by capturing the context of global information. [38] attains better performance by learn a nonlinear hyperplane to capture more discriminative features. [13] outperforms the accuracy of state-of-the-art by improving dense trajectory that based on bag of word, fisher vector, and high abstraction features being extracted from spatio-temporal data with I3D model. However, this method trades the performance with the complexity due to highly algorithm, this utilizes combined CNN features of RGB and optical flow materials, then transform to 4 types feature descriptor streams of bag of words, fisher vector first order, fisher vector second order, and high abstraction feature. The output of 4 types of feature descriptors is combined and fed into the prediction network for classification. Our proposed method, which utilized I3D and attention, providing promising performance that can be compared to state-of-the-art as our method used attention module to capture learn temporal dependencies among the frames in the videos.

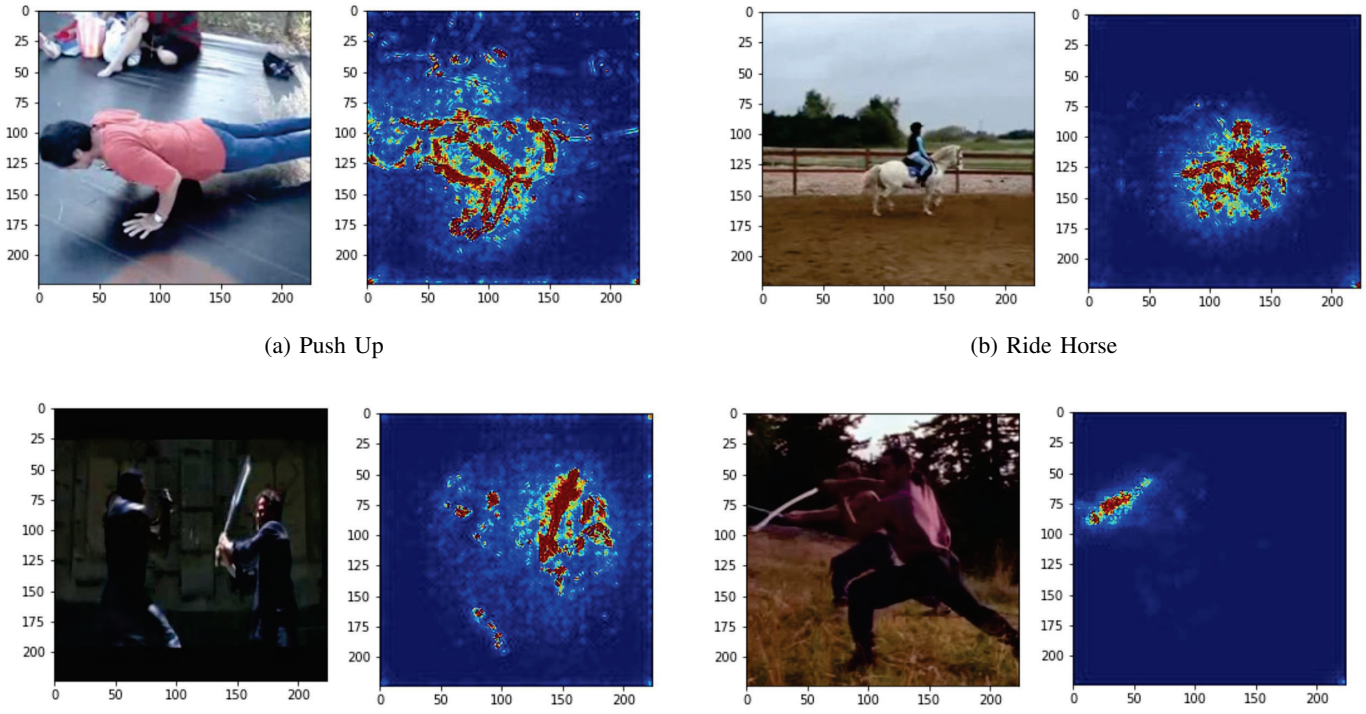
V. Conclusion

This study investigates action recognition in video clips. We propose a two-stream 3D convolution attentional network, which is able to capture the long-term dependency of 3D CNN features, that is essential in recognizing different actions. The proposed method combines the spatio-temporal features of RGB and the optical flows of images. Using the popular HMDB-51 dataset, the ablation studies show that our model can provide a promising performance with less complex model.

References

[1] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[2] M. Ryoo, B. Rothrock, C. Fleming, and H. Yang, "Privacy-preserving human activity recognition from extreme low resolution." in *Proceedings*



(a) Push Up

(b) Ride Horse

(c) Sword

(d) Sword Exercise

Fig. 4: Feature visualization on (a) and (b) images are the best recognized action, (c) and (d) images are the worst recognized action.

of the Association for the Advancement of Artificial Intelligence, 2017, pp. 4255–4262.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[4] M. Quigley, K. Mohta, S. S. Shivakumar, M. Watterson, Y. Mulgaonkar, M. Arguedas, K. Sun, S. Liu, B. Pfrommer, V. Kumar *et al.*, "The open vision computer: An integrated sensing and compute system for mobile robots," in *2019 International Conference on Robotics and Automation*. IEEE, 2019, pp. 1834–1840.

[5] R. R. A. Pramono, Y.-T. Chen, and W.-H. Fang, "Hierarchical self-attention network for action localization in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 61–70.

[6] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1647–1656.

[7] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "Mars: Motion-augmented rgb stream for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7882–7891.

[8] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "Potion: Pose motion representation for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7024–7033.

[9] A. Piergiovanni and M. S. Ryoo, "Representation flow for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9945–9953.

[10] D. Purwanto, Y.-T. Chen, and W.-H. Fang, "Temporal aggregation for first-person action recognition using hilbert-huang transform," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 895–900.

[11] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[13] L. Wang, P. Koniusz, and D. Q. Huynh, "Hallucinating bag-of-words and fisher vector idt terms for cnn-based action recognition," *arXiv preprint arXiv:1906.05910*, 2019.

[14] I. Laptev, C. Marszalek, M.and Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[15] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proceedings of British Machine Vision Conference*, 2008, pp. 275–1.

[16] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.

[17] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2009, pp. 58–65.

[18] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1234–1241.

[19] F. Baumann, A. Ehlers, B. Rosenhahn, and J. Liao, "Recognizing human actions using novel space-time volume binary patterns," *Neurocomputing*, vol. 173, pp. 54–63, 2016.

[20] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

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

[22] A. Noguchi and K. Yanai, "A surf-based spatio-temporal feature for feature-fusion-based action recognition," in *European Conference on Computer Vision*. Springer, 2010, pp. 153–167.

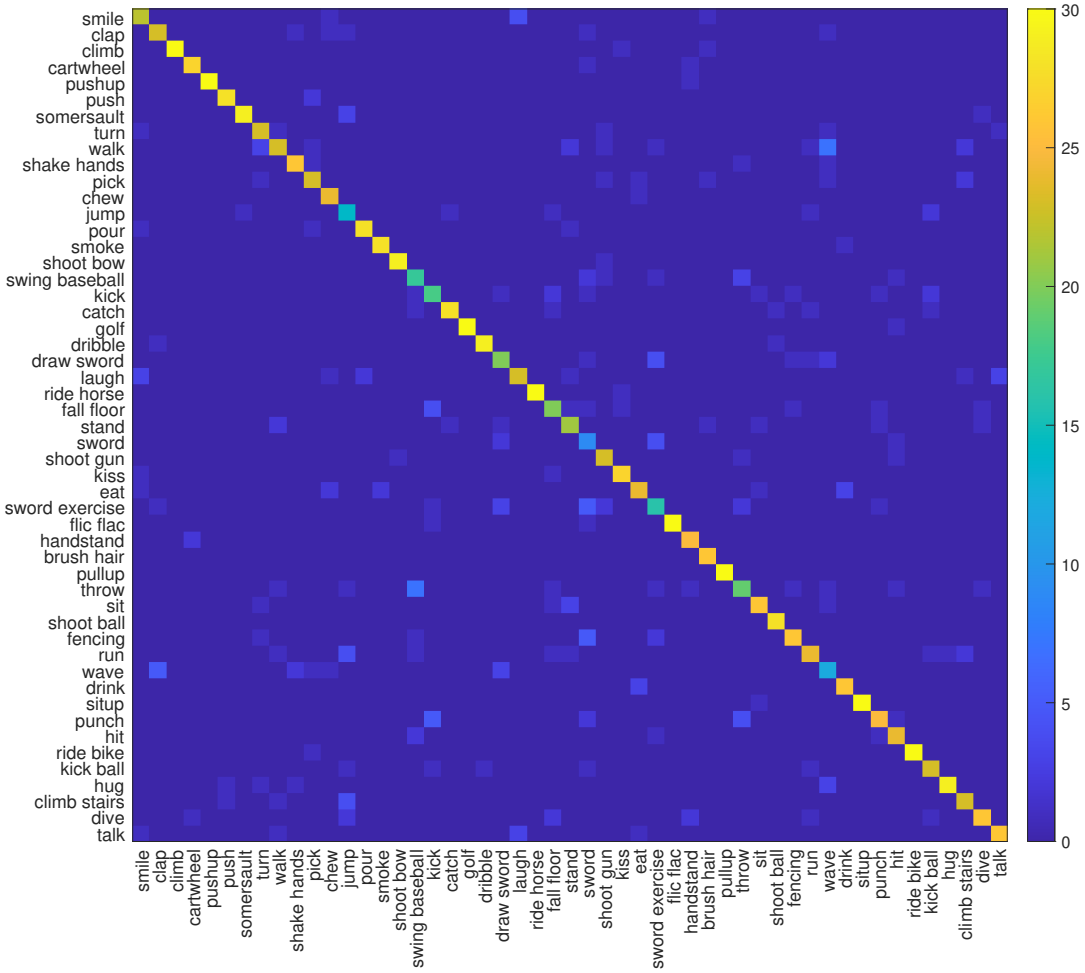


Fig. 5: Confusion matrix of each class of HMDB dataset.

[23] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 436–450, 2012.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4305–4314.

[26] 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.

[27] T. Shen, T. Zhou, G. Long, J. Jiang, and C. Zhang, "Bi-directional block self-attention for fast and memory-efficient sequence modeling," in *Proceedings of the International Conference on Learning Representations*, 2018.

[28] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," *arXiv preprint arXiv:1704.02971*, 2017.

[29] Y. Wang, S. Wang, J. Tang, N. O'Hare, Y. Chang, and B. Li, "Hierarchical attention network for action recognition in videos," *arXiv preprint arXiv:1607.06416*, 2016.

[30] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[31] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[33] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," *Pattern Recognition*, pp. 214–223, 2007.

[34] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[35] K. W. Cheng, Y. T. Chen, and W. H. Fang, "Improved object detection with iterative localization refinement in convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2017.

[36] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2556–2563.

[37] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[38] J. Wang, A. Cherian, F. Porikli, and S. Gould, "Video representation learning using discriminative pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1149–1158.