EasyChair Preprint

№ 10797

A Temporal Difference Pyramid Network for Action Recognition

Liu Linfu, Yun Tie, Qi Lin and Chengwu Liang

August 29, 2023

TDPN:A Temporal Difference Pyramid Network for Action Recognition

***Abstract*—The visual rhythm of human actions can distinguish human actions with high visual similarity by expressing the dynamic and rhythmic scales of activities. In traditional convolutional neural networks, most of the input videos are sampled at different rates using local receptive fields, and there are also methods using multi-layer networks to process videos at different rates, but this requires high computational costs and additional computational resources. The previous methods cannot deal with the problem of feature fusion between different levels of features when identifying the relationship between visual speed and high-level attributes and fine-grained rhythm. In our study, we propose a Temporal Difference Pyramid Network (TDPN). What is significantly extracted from the action recognition network is multi-layer backbone features. The two key points are the global temporal modeling module and the local temporal modeling module. The global temporal modeling module is responsible for extracting low-level features with more location information, while the local temporal modeling module is responsible for extracting high-level features with more semantic information.Meanwhile, a multi-head attention mechanism has been used to simultaneously focus on different levels of feature information, better capturing the relative relationships between different features in the video sequence.Then, by aggregating different levels of features, the pixel-level fine-grained rhythm dynamics of action visual rhythm with rich information is obtained.Tests on various action recognition benchmarks, including Something Something V1 and V2 and Kinetics-400 , have shown that the TDPN network we suggested significantly boosts the performance of the current video-based action recognition model.**

***Index Terms*—Action Recognition, Visual tempo, TDPN , pyramid network ,Temporal Difference**

I. Introduction

In the realm of computer vision, the recognition of actions in the video is a prominent topic for research. The primary goal is to automatically identify human activity in the video feed. With the widespread use of convolutional neural networks, circulating neural networks, and deep learning methods The field of action recognition rapidly incorporates deep learning methods. These techniques essentially decompose video data into a collection of vector sequences before extracting the features. The spatial and angular sequence of motion is represented by these vectors. The identified actions are sorted out by High way of picture or video features.

In action recognition, the study of dynamic visual tempo can help us better understand videos and images and analyze human behavior and movements. Therefore, research on the visual tempo of the action instance that characterizes different actions can more deeply understand the characteristics of different types of actions and its impact on the action recognition. As shown in Figure 1, the speed of the mobile object in Figure 1 (a) is faster than the speed in Figure 1 (b). In Figure 1 (b), the movement of the mobile object is very subtle, but in Figure 1 (a), because the speed of the mobile object is particularly fast, only the latter frame images show the movement of the mobile object. Show the movement of the mobile object. At the same tempo, the angle of the mobile objects taken by the camera in the two pictures is also different. These are caused by changes in the visual tempo of the action. Figure 1 shows that there are significant differences between the variation coefficients between different categories, such as wearing headphones and turning their heads, there is a large visual difference. Therefore, how to accurately identify the visual tempo of similar actions has a vital role in improving the accuracy of action recognition.The action examples above show that people tend to act at different tempos even for the same action. The plot below shows different action categories sorted by their variances of visual tempos. In earlier research, Slow fast[6] created a dual-channel

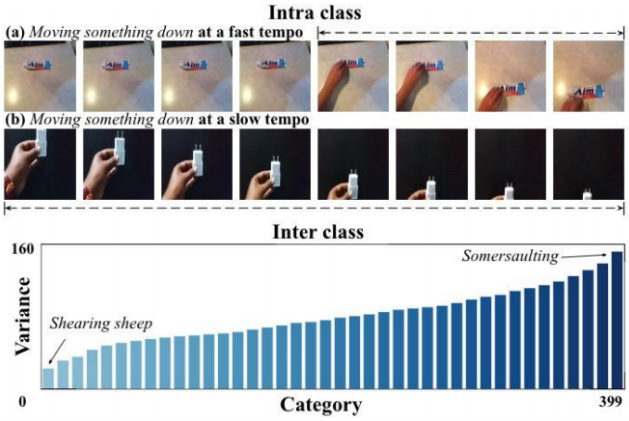

Fig. 1. Visual tempo variation of intra- and inter-class.

Slow fast model for video identification by sampling video frames at two different rates as input operating at a frame rate. The primary network aggregates fast-paced and slow-paced data before handling the action instance using two different tempo standards,The dynamic visual tempo of the input frame level processing instance is still expensive to calculate because this method requires variable frame sampling rates. A trained nerve module is offered by Motion Squeeze[4] for determining a correspondence between frames and transforming it into motion characteristics. These techniques offer the potential to finely simulate the tempo dynamics of subsequent frames, but

they disregard the significance of the action's visual tempo.In this effort, our contribution to resolving the aforementioned issues is as follows:

- We proposed the global and local temporal modeling modules are two modules , which first introduced temporal difference operations into the pyramid network, and through multi-level feature extraction and fusion, we improved the accuracy of action feature recognition.
- In the Temporal Difference Pyramid Network, we introduced a multi-head attention mechanism, where each head can learn a different attention weight distribution. This allows the model to simultaneously focus on different levels of feature information, better capturing the relative relationships between different features in the video sequence.
- Extensive research conducted on various commonly used action recognition datasets shows that the proposed network model (TDPN) significantly improves the performance of current video-based action recognition models.

II. Related Work

A. Video action recognition research based on deep learning

Deep learning approaches can be categorized into two groups: 2D convolutional networks and 3D convolutional networks. Convolutional neural networks (CNNs) are used in 2D CNN to extract the characteristic of video frames. The size of the feature mapping is gradually lowered and more abstract feature information is recovered by coupling numerous convolutional layers and pooling layers. The Two-Stream[5] method, put out by Simonyan and Zisserman and others, models the space and timing characteristic of the video using two independent CNN networks before categorizing the output outcomes of the two CNNs. TSN[3] combines the 2D CNN network with the LSTM structure to categorize video motions. [7, 8]was proposed by Wang and colleagues., The video frame sequence is broken up into several sub-sequences, and the 2D CNN network is used to extract the spatial features of each new sequence and classify the global characteristics of every subsequence. In order to effectively capture the action characteristics in the video, 3D CNN can take into account tempo and location information simultaneously. "Tran et.al" Introduced the C3D[9] convolutional neural network, which uses a tempo convolutional layer to collect the temporal information of the video 3D convolution is used[13] in the space and tempo dimensions to extract more thorough space tempo characteristic. A 3D CNN model called RES3D[11] is built on the res-net architecture, by include residual connections and 3D convolution processes, this model improves the extraction of space-tempo features. A novel 3D CNN model called Slow-fast Networks combines two CNN branches operating at various rates to process fast and slow motion in the video[20]. Another 3D CNN model built on the Res-net architecture is R (2+1) D [10]. To extract space-tempo features more effectively, it combines 1D convolution and 2D convolution.

B. Visual tempo construction model in action recognition

Action tempo detection was made difficult by the intricate structures of action instances, particularly when many visual tempos were combined. Slow-fast hard-codes the variations in visual tempo using the input-level frame pyramid,the pyramid network extract feature at varying rates level by level. Additionally, a network that incorporates the intermediate feature of these networks separates each layer of the pyramid, pyramids network's frame and specialized networks enable to handle Slow-fast variations in visual tempo gradually. In order to create feature pyramids that may enter any length of input videos, DTPN is additionally sampled to frames at various frames/seconds (FPS) rates.

For input videos of any length, DTPN captures the video frames at various frame sampling rates and expresses pyramid characteristics that can describe both slow and fast-paced tempo dynamics. When frame sampling rates rise, this sampling approach frequently necessitates numerous frames, which results in high calculation costs. A two-stage frame pyramid is used by Slow-fast to hard-codes the visual tempo of the visual tempo[18]. The feature properties of these branches are interactive, and the branches are carefully constructed to process each level separately. Although the multi-branch network is expensive, it can gradually handle changes in the visual tempo of action. But cannot use low-level characteristics, and the useful long-range fine-grained tempo dynamic from long-range frames from the high-level characteristics obtained by overlapping multiple local tempo convolution may be weakened[19].

C. Temporal difference algorithm

Temporal Difference (TD) algorithm is a reinforcement learning technique used to estimate the value function of a Markov Decision Process (MDP) through a combination of bootstrapping and sampling. The basic idea behind the Temporal Difference algorithm is to update the estimated value of a state based on the difference between the observed reward and the estimated value of the next state. The algorithm learns by iteratively updating the value function based on the observed rewards and the predictions made by the current value function.The algorithm works as follows:

- Initialize the value function: Start with an initial estimate of the value function for each state in the MDP.
- Interaction with the environment: Agent takes actions in the environment based on its current policy and observes the next state and the reward.
- Update the value function: Use the observed reward and the estimated value of the next state to update the value of the current state. The update is performed by adjusting the estimated value towards the target value, which is the sum of the observed reward and the estimated value of the next state discounted by a factor called the discount rate.
- Repeat steps 2 and 3: Continue interacting with the environment, updating the value function, and moving to

the next state until the desired convergence is reached or a predefined number of iterations is completed.

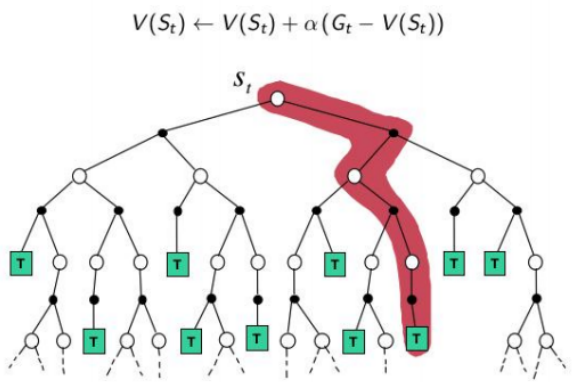

Fig. 2. Difference principle of time difference algorithm.

The Temporal Difference algorithm combines ideas from both dynamic programming and Monte Carlo methods. It uses bootstrapping to estimate the value function by updating values based on estimates of future states. This allows TD algorithms to learn online, incrementally update the value function after each interaction with the environment, and handle large state spaces efficiently.One of the most well-known Temporal Difference algorithms is Q-learning, which uses a similar update rule to estimate the action-value function (Q-function) instead of the state-value function.The TD algorithm are highly practical in real-world applications and can handle difficult situations like continuous state and action space. In order to induce changes in the transportation of the sequential information, our work updated the backbone layers in the pyramid structure of the backbone network and joined the straightforward operation of Difference. The feature fusion is then carried out following the spatial semantic adjustment and tempo rate adjustment.

D. Attention mechanism

Attention mechanism plays an important role in deep learning, which was first introduced into the field of natural language processing to enhance the attention of neural networks to different positions or elements in input sequences. Subsequently, attention mechanisms were widely applied in fields such as computer vision and speech recognition.The attention mechanism can be seen as a method of selecting and weighting information, enabling the model to allocate attention resources targeted to different parts of the input. By considering the correlation and importance of inputs, attention mechanisms can effectively extract key information and ignore irrelevant or noisy content.In the field of natural language processing, the common attention mechanism is the Self Attention mechanism[33] proposed based on the Transformer model. The self attention mechanism allows the model to weight and focus on different positions of the input sequence when generating each word, thereby better capturing contextual information.

In the field of computer vision, attention mechanisms are widely used in tasks such as image classification, object detection, and image generation. The common attention mechanism is the spatial attention mechanism[34], which allows models to weight pixels at different spatial positions when processing images. In this way, the model can focus on the most important or distinctive areas in the image, improving the performance and accuracy of the task.Attention mechanism provides an effective way to select and weight different parts of information in image processing, enabling models to more accurately understand and process input data. The introduction of attention mechanisms has brought significant performance improvements to many tasks and has become a highly focused research direction in the field of deep learning.

III. Our Work

The alignment and correctness of the spatial and semantic information of the image will have a significant impact on the action recognition, because the visual rhythm recognizes the dynamic and rhythm scale of the action. TPN [12] constructs a visual rhythm pyramid by compiling the main features of each layer and aggregating them to capture the visual rhythm data of feature-level actions. This method is inspired by the feature-level pyramid network [14] and can handle significant differences in spatial scales. TPN has been continuously improved on various action recognition datasets, but the gain advantage of low-level features in action recognition is limited due to the ability to rely on the main network for rhythm modeling. Also TPN does not make full use of the relevant fine-grained rhythm dynamics and scale information, because the features extracted by the high-level network contain more semantic information, and the features extracted by the low-level network can provide a wide range of fine-grained rhythm dynamics and rhythm scale information. It is unwise to abandon the low-level action visual rhythm features while extracting the high-level action visual rhythm features to improve performance.

The time derivative [16] is highly correlated with the optical flow and can be used to describe the speed or acceleration of the pixel value over time to capture the dynamic features in the image sequence. The common time derivative operator is the Horn-Schunck algorithm in the optical flow algorithm [26], which estimates the motion speed of pixels based on the optical flow constraint equation. Applying the reciprocal of time to the action recognition of video can accurately model the internal and inter-class differences of the visual rhythm of the action instance, which may bring significant improvement to the action recognition.

A. Backbone of our module

The backbone network that we employ is 3D Res-net. The backbone network of TDPN is constructed using the output properties of res2, res3, res3, and res5, as depicted in Table 1. They have each reduced the space by 4, 8 and 16 in comparison to the input frames. 32 tempos as well.

We make a simple modification to the backbone network by replacing the spatial convolutional layers with the multi-head

self-attention (MHSA) layers proposed in Transformer(As Fig3 shown). MHSA allows the model to learn different relations separately on different attention heads. This enables our model to simultaneously capture multiple different network layers and different types of semantic relationships, leading to a better understanding of the information in the input data. At the same time, the attention heads of MHSA are independent of each other, which allows it to perform parallel computing on multiple heads at the same time, which helps to improve the training and inference efficiency of the model. More importantly, MHSA can better model long-distance semantic connections by learning multiple different dependencies, and finally feature representations at different levels.

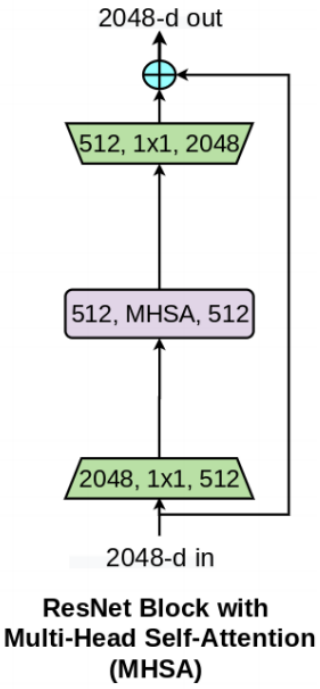

Fig. 3. An instantiation of the Bottleneck Transformer as a ResNet bottleneck block.

As we can see, our network backbone model incorporates replace space 3×3 convolutional layers with multi head self attention (MHSA).

B. Local temporal modeling module

In The backbone network that we employ is 3D Res-net. The backbone network of TDPN is constructed using the output properties of res2, res3, res3, and res5, as depicted in Table 1. They have each reduced the space by 4, 8, 16 and 32 tempos as well in comparison to the input frames .In order to processing videos successful and properly encode the appearance and sports information, our local temporal modeling module chooses one RGB frame with a tempo difference, a single-frame RGB may detect local movements by using fuzzy tempo difference information for low-level feature extraction[23]. According to Figure 4, the local temporal modeling module, after sampling the input video frames, we

TABLE I
BACKBONE OF OUR RES-NET

Stage	Layer	Output size
raw	-	$8 \times 224 \times 224$
Conv 1	$1 \times 7 \times 7, 64$,stride 1	$8 \times 112 \times 112$
Conv 2	$1 \times 7 \times 7, 64$, stride $1, 2, 2$	$8 \times 112 \times 112$
Pool 1	$1 \times 3 \times 3$,mm ,stride $1, 2, 3$	$8 \times 56 \times 56$
Res2	$\begin{bmatrix} 1 \times 1 \times 1, 64 \\ 1 \times 3 \times 3, 128 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 3$	$8 \times 56 \times 56$
Res3	$\begin{bmatrix} 1 \times 1 \times 1, 128 \\ 1 \times 3 \times 3, 256 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 4$	$8 \times 28 \times 28$
Res4	$\begin{bmatrix} 1 \times 1 \times 1, 256 \\ 1 \times 3 \times 3, 256 \\ 1 \times 1 \times 1, 1024 \end{bmatrix} \times 3$	$8 \times 14 \times 14$
Res5	$\begin{bmatrix} 1 \times 1 \times 1, 512 \\ MHSA, 512 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 3$	$8 \times 7 \times 7$
Global average pool,fc		$1 \times 1 \times 1$

perform differentiated operations on them before adding the channels of the subsequent tempo frames and finally adding the channels of various dimensions. Utilize a 2D CNN to extract the motion-related feature[27], then extract the features, and then use a horizontal connection to match the original RGB features. Adding our local temporal module to the high level of the pyramid network can effectively make up for the capture of fine particle size features in the low -level network of the pyramid network to reduce the loss of characteristics.

$$\mathcal{H}\left(I_{\mathrm{i}}\right) = \text{ Upsmp } \left(\text{CONV}\left(\text{ Downsmp } \left(\mathbf{D}\left(I_{\mathrm{i}}\right)\right)\right)\right). \quad (1)$$

$\mathcal{H}$ denotes our local temporal modeling module. D indicates that cent RGB of i indicates that CNN is a lightweight module, which is used to operate stacked RGB differences, and follows low resolution processing strategies.

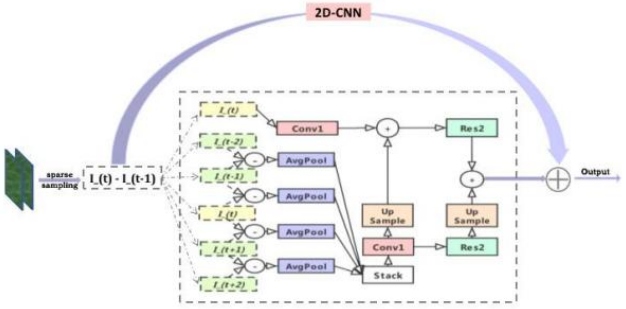

Fig. 4. Local temporal modeling module backup.

C. Global temporal modeling module

Due to the difficulty in aligning the spatial location between the low-level video frames of the pyramid network. We employ a differential module with a low resolution and light weight for the global temporal modeling module. Utilize two-way differential modules with multiple scales to track cross-segment variations in movement motivation[24]. The unique two-way

and multi-scale velocity differential module in the global temporal modeling module enhances the original reproduction by using cross-segment information. Through a multi-scale architecture, the difference in the differential computation is smooth. The adjacent fragment's adjacent fragment's tempo difference is as follows:

$$\mathrm{C}\left(F_i, F_{i+1}\right) = F_i - \mathrm{Conv}\left(F_{i+1}\right) \tag{2}$$

Among them, C represents the differences between alignment between F in different level, and conv is to achieve smooth channel convolution.

After the temporal difference between alignment is smooth, the two-way span for temporal difference is lastly employed to improve the frame level characteristics, followed by the multi-scale module for remote exercise information extraction:

$$\begin{aligned} F_i \odot \mathcal{G}\left(F_i, F_{i+1}\right) = F_i \odot \frac{1}{2}\left[M\left(F_i, F_{i+1}\right)\right. \\ \left. + M\left(F_{i+1}, F_i\right)\right] \end{aligned} \tag{3}$$

Where $\odot$ is the element-wise multiplication. We also combine the original frame level representation and enhance representation via a residual connection.$\mathcal{G}$ denotes our global temporal modeling module.

We let the global temporal modeling module extracts features in the low-level network, while the local temporal modeling module extracts features in the high-level network. Therefore, the operation of the motion information carried when operating top-down and bottom-up in the pyramid structure is more abundant [25]. By flowing down from the top, the features extracted by the low-level network can be extracted by the high-level network to improve discriminativeness. The accuracy rate will be greatly improved. Bottom-up flow [28], the general features of the global context are fused with the detailed features of the rich local context, which can make full use of the global and local information of the trajectory.

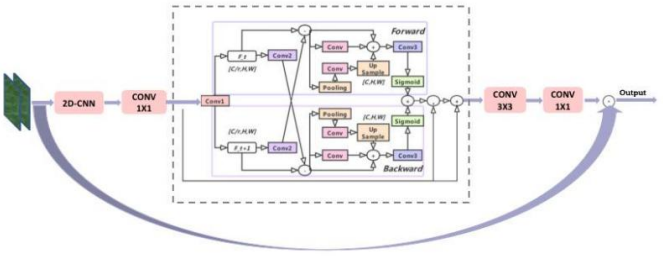

Fig. 5. Local temporal modeling module backup.

IV. Experiments

A. Datasets

We were prepared to evaluate our network model on the two video datasets, and these two data sets are concerned about the different aspects of action examples used for identification. Kinetics-400[32] is a large YouTube video data-set with about 300,000 pruning videos, covering 400 categories, the dynamic data-set contains activities in daily life, and some categories are highly related to interactive objects or scene context. We train our TDPN based on training data (240K video) and report performance based on verification data (20K video). Something Something[30] is a large -scale data set created by crowd sourcing. The video is collected by forming the same action by different objects, so the action recognition should pay attention to the movement features, not the object or scene context. The first version includes about 100,000 videos in more than 174 categories, and the second version contains more videos. In training concentration, it contains about 169K videos. The verification set contains about 25K video.

We utilize the Image Net data-set to train our TDPN beforehand. The initial learning rate is 0.02 and the batch size is 128. The number of training cycles is set in the data set to 100 and in other data sets to 60. The learning rate will be divided by a coefficient of 10 when the saturation rate is determined by the verification performance. Each video's short-sized size was changed to 256 for testing purpose.

Specifically, a 2D kernel of size k×k will be inflated to have the size t×k×k, with its original weights copied for t tempos and rescaled by 1/t. Following the setting in[5], the input frames are sampled from a set of consecutive 64 frames at a specific interval τ .The augmentation of horizontal flip and a dropout of 0.5 are adopted toreduce over fitting. And Batch Norm (BN) is not frozen.We use a momentum of 0.9, a weight decay of 0.0001 and a synchronized SGD training over 6 GPUs. Each GPU has a batch-size of 8, resulting in a mini-batch of 64 in total. For Kinetics-400, the learning rate is 0.001 and will be reduced by a factor of 10 at 85, 105 epochs (120 epochs in total)respectively. For Something-Something V1 and V2 [10], our model is trained for 120 and 60 epochs separately.

B. Results

TABLE II
COMPARISON WITH OTHER STATE-OF-THE-ART METHODS ON THE VALIDATION SET OF KINETICS-400

Model	Layer	Flow	Top-1	Top-5
Two-Stream I3D[4]	64	√	75.7	92
SlowFast-R50[15]	64	√	76.5	92.6
TSN-R101[17]	32		76.8	92.1
TPN-R50[12]	32 × 2		77.7	93.3
TPN-R101[12]	32 × 2		78.9	93.5
AGPN[36]	32 × 2		79.0	94.0
TDPN-R50(ours)	32 × 2		78.9	93.8
TDPN-R101(ours)	32 × 2		**79.6**	**94.3**

According to the experimental results on the validation set of Kinetics-400 in Table 2, it can be seen that different networks achieve different accuracies. Our TDPN-R50 can achieve an accuracy of 0.789, while the same TDPN-R101 can achieve an accuracy of 0.796. Compared to the original method, both methods have been improved.On the same network foundation, our TDPN achieves better accuracy.

According to the experimental results on the validation set of Something Something V1 and V2. in Table 3, it can be

seen that different networks achieve different accuracies. Our TDPN-R50 can achieve an accuracy of 0.783, while the same TDPN-R101 can achieve an accuracy of 0.792. Compared to the original method, both methods have been improved.On the same network foundation, our TDPN achieves better accuracy.

TABLE III
COMPARISON WITH OTHER STATE-OF-THE-ART METHODS ON THE VALIDATION SET OF KINETICS-400

Model	Layer	Flow	Top-1	Top-5
Two-Stream I3D[4]	64	√	75.3	91.8
SlowFast-R50[15]	64	√	76.5	92.6
TSN-R101[17]	32		76.7	92.1
TPN-R50[12]	32 × 2		77.7	93.2
TPN-R101[12]	32 × 2		77.9	93.3
AGPN[36]	32 × 2		79.0	93.6
TDPN-R50(ours)	32 × 2		78.3	93.5
TDPN-R101(ours)	32 × 2		**79.2**	**93.8**

C. Comparative test

To understand TDPN better, we evaluate our network with different variants on our data-sets to study their effects. For all experiments, we only change a certain part of our model and use the same evaluation settings.

TABLE IV
ABLATION EXPERIMENTS OF TWO TEMPORAL MODELING MODULES.

None	Global in low-level	Local in high-level	Top-1
√			74.0
	√		75.8
		√	75.7
	√	√	**77.2**

We conduct ablation experiments on the two temporal modeling modules on the related model, and explore the recognition accuracy of the pyramid network when combined with the global temporal modeling module and the local temporal modeling module, respectively. According to Table 4, we get the best results when letting the global temporal modeling module extracts features in the low-level network, while the local temporal modeling module extracts features in the high-level network.

TABLE V
COMPARING THE RESULTS OF EQUIPPING OUR TDPN WITH OTHER METHODS

Model	Frames	NOUN@1	VERB@1
TSN(Flow)[3]	25	28.5	45.5
TSN(Fusion)[3]	25	36.5	47.6
TSN+TPN	8	41.3	58.6
AGPN[36]	8	42.4	61.1
TSN+TDPN(ours)	8	**43.2**	**62.5**

The results in Table 5 shows that when used in a plug-and-play way with other models, our suggested TDPN network performs better than the original approach.

V. CONCLUSIONS

This paper proposes the Temporal Difference Pyramid Network (TDPN). It can be embedded into some important action recognition network models in a plug-and-play manner, and can accurately extract action recognition models with multi-layer backbone network features. The global temporal modeling module and the local temporal modeling module are our main innovations. The global time modeling module is responsible for extracting low-level network features with more position information, while the local velocity modeling module is responsible for extracting more semantic information in high-level network. High-level features are then aggregated across different levels to obtain pixel-level network fine-grained rhythmic dynamics for action-visual rhythms that contain rich information. Meanwhile, a multi-head attention mechanism has been used to simultaneously focus on different levels of feature information, better capturing the relative relationships between different features in the video sequence.Finally, our experiments on multiple action recognition benchmarks such as Something SomethingV1and V2, Kinetics-400 etc. demonstrate that our proposed TDPN greatly improves the performance of existing video-based action recognition models.

REFERENCES

[1] Chauhan R, Ghanshala K K, Joshi R C Convolutional neural network (CNN)for image detection and recognition[C]//2018 first international conference on secure cyber computing and communication (ICSCCC).IEEE.(2018)

[2] Medsker L R, Jain L C.Recurrent neural networks[J]. Design and Applications.(2011)

[3] L. Wanget al.,"Temporal segment networks for action recognition in videos,"IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 11,pp. 2740-2755.(2019)

[4] H. Kwon, M. Kim, S. Kwak, and M. Cho,"Motion squeeze: Neural motion feature learning for video understanding," inProc. Eur.Conf.Comput. Vis,pp.345-362.(2020)

[5] Simonyan and A. Zisserman,"Two-stream convolutional networks for action recognition in videos, " in Proc. Int. Conf. Neural Inf. Process.Syst.pp.568-576.(2014)

[6] Feichtenhofer C, Fan H, Malik J, et al.Slow fast networks for video recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision.6202-6211.(2019)

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri,"Learning spatio-temporal features with 3Dconvolutional networks," in Proc. IEEE Int. Conf. Comput.Vis. (ICCV), Dec.,pp. 44894497.(2018)

[8] K. Hara, H. Kataoka, and Y. Satoh,"Can spatio-temporal 3D CNN sretrace the history of 2D CNNs and Image Net?" inProc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. pp. 6546-6555.(2018)

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri,"Learning spatio-temporal features with 3Dconvolutional networks," inProc. IEEE Int. Conf. Comput.Vis.(ICCV), Dec.pp.4489-4497.(2015)

[10] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann Le Cun, and Manohar Paluri.A closer look at spatio-temporal convolutions for action recognition.InProc.CVPR.(2018)

[11] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M: A closer look at spatio-temporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.pp.6450-6459 (2020)

[12] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou,"Temporal pyramid network for action recognition," inProc. IEEE Conf. Comput.Vis.Pattern Recognit.,Jun.pp.591-600.(2020)

[13] Wang Y, Song J, Wang L, et al.Two-Stream 3D-CNNs for Action Recognition in Videos[C]//Bmvc.(2020)

[14] Zhang, X. Dai, and Y.-F. Wang,"Dynamic temporal pyramid network:A closer look at multi-scale modeling for activity detection, "(2018) inProc.Asian Conf.Comput.Vis.pp.712-728.

[15] X. Wang, R. Girshick, A. Gupta, and K. He"Non-local neural networks, "inProc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,Jun. pp. 7794-7803.,(2018)

[16] Liu Q, Sung A H, Qiao M.Temporal derivative-based spectrum and mel-cepstrum audio steg analysis[J]. IEEE Transactions on Information Forensics and Security, 4(3): 359-368.(2019)

[17] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kai ming He,Bharath Hariharan, and Serge Belongie.Feature pyramid networks for object detection.InProc.CVPR.(2017)

[18] J.Piergiovanni and M.S.Ryoo,"Representation flow for action recognition,"inProc. IEEE Conf. Comput. Vis. Pattern Recognit.Jun.pp. 9945-9953.(2019)

[19] Y. Ji, Y. Zhan, Y. Yang, X. Xu, F. Shen, and H. T. Shen," A context knowledge map guided Coarse-to-Fine action recognition, " IEEE Trans.Image Process.vol.29,pp.2742-2752.(2021)

[20] B.Zhou,A.Andonian, A.Oliva and A. Torralba," Temporal relational reasoning in videos," inProc.Eur.Conf.Comput.Vis,pp.803-818.(2018)

[21] Lin,C.Gan,and S. Han,"TSM: Temporal shift module for efficient video understanding," inProc.IEEE Int.Conf.Comput.Vis.Oct.pp.(2021)

[22] Zheng, Z. Liu, T. Lu, and L.Wang,"Dynamic sampling networks for efficient action recognition in videos, " IEEE Trans. Image Process.,vol. 29,pp.7970 -7983.(2022)

[23] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri,"A closer look at spatio-temporal convolutions for action recognition, " inProc. IEEE Conf.Comput. Vis.Pattern Recognit.,Jun.pp.6450-6459.(2021)

[24] Z.Tuetal."Action-stage emphasized spatio-temporal VLAD for video action recognition,"IEEE Trans. Image Process., vol.28, no.6.pp. 2799-2812,Jan.(2019)

[25] S.-Z.Wang,Y.-S.Chen, S.-H. Lee, and C.-C. J.Kuo,"Visual tempo analysis for MTV-style home video authoring," inProc. Congr. Image Signal Process.vol.2,May ,pp.192-196.(2018)

[26] M. Zolfaghari, K. Singh, and T. Brox,"ECO: Efficient convolutional network for online video understanding, " inProc. Eur. Conf. Comput.Vis. Sep.pp.695-712.(2021)

[27] X. Li, B. Shuai, and J. Tighe," Directional temporal modeling for action recognition," inProc.Eur.Conf.Comput.Vis.(2022)

[28] Liu, L. Wang, W.Wu, C. Qian, and T. Lu,"TAM: Temporal adaptive module for video recognition," .(2021)

[29] L. Wang, Z. Tong, B. Ji, and G. Wu,"TDN: Temporal difference networks for efficient action recognition," inProc. IEEE/CVF Conf.Comput. Vis. Pattern Recognit.(CVPR).(2021)

[30] Da Zhang, Xiyang Dai, and Yuan-Fang Wang. Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection. In Asian Conference on Computer Vision, pages 712–728, 2018.

[31] Goyalet al."The 'something something' video database for learning and evaluating visual common sense," inProc. IEEE Int. Conf.Comput.Vis.(2019)

[32] W. Kayet al."The kinetics human action video dataset".(2019)

[33] Guo M H, Xu T X, Liu J J, et al.Attention mechanisms in computer vision: A survey[J]. Computational visual media,8(3): 331-368.(2022)

[34] Rizzolatti G, Craighero L.Spatial attention: Mechanisms and theories[J]. Advances in psychological science,, 2: 171-198.(1998)

[35] Geng, Tiantian, et al. "Spatial-Temporal Pyramid Graph Reasoning for Action Recognition." IEEE Transactions on Image Processing 31 (2022): 5484-5497.

[36] Chen, Yatong, et al. "Agpn: Action granularity pyramid network for video action recognition." IEEE Transactions on Circuits and Systems for Video Technology (2023).