



Learning Visual Tempo for Action Recognition

Mu Nie¹, Sen Yang², and Wankou Yang²(✉)

¹ School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China

² School of Automation, Southeast University, Nanjing 210096, China

wkyang@seu.edu.cn

Abstract. The variation of visual tempo, which is an essential feature in action recognition, characterizes the spatiotemporal scale of the action and the dynamics. Existing models usually use spatiotemporal convolution to understand spatiotemporal scenarios. However, they cannot cope with the difference in the visual tempo changes, due to the limited view of temporal and spatial dimensions. To address these issues, we propose a multi-receptive field spatiotemporal (MRF-ST) network in this paper, to effectively model the spatial and temporal information. We utilize dilated convolutions to obtain different receptive fields and design dynamic weighting with different dilation rates based on the attention mechanism. In the proposed network, the MRF-ST network can directly obtain various tempos in the same network layer without any additional learning cost. Moreover, the network can improve the accuracy of action recognition by learning more visual tempo of different actions. Extensive evaluations show that MRF-ST reaches the state-of-the-art on the UCF-101 and HMDB-51 datasets. Further analysis also indicates that MRF-ST can significantly improve the performance at the scenes with large variances in visual tempo.

Keywords: Action recognition · Spatiotemporal · Multi-receptive field · Visual tempo

1 Introduction

Video is the lifeblood of the internet, which means analyzing and understanding video content is critical for the most modern artificial intelligence agents [1, 2]. Deep neural networks play an important role in many aspects [3, 4]. Although the accuracy of video action recognition has been greatly improved [5, 6], in the design of these recognition networks, an important aspect that characterizes different actions - the visual tempo of action instances is often overlooked. Unfortunately, existing models [7, 8] mainly focus on using spatiotemporal factorization to reduce computational cost and model parameters.

Visual tempo actually describes the speed at which an action is performed, which often determines the effective temporal of recognition. Therefore, we need to consider the differences in the temporal and spatial feature of action instances, when designing an action recognition network. For example, walking action is slower than running class in temporal and spatial change frequency. Action tempo not only exists inter-class actions, but also has significant differences in the intra-class. In Fig. 1, we show examples of the video clips and coefficients of variations from the HMDB-51 dataset.

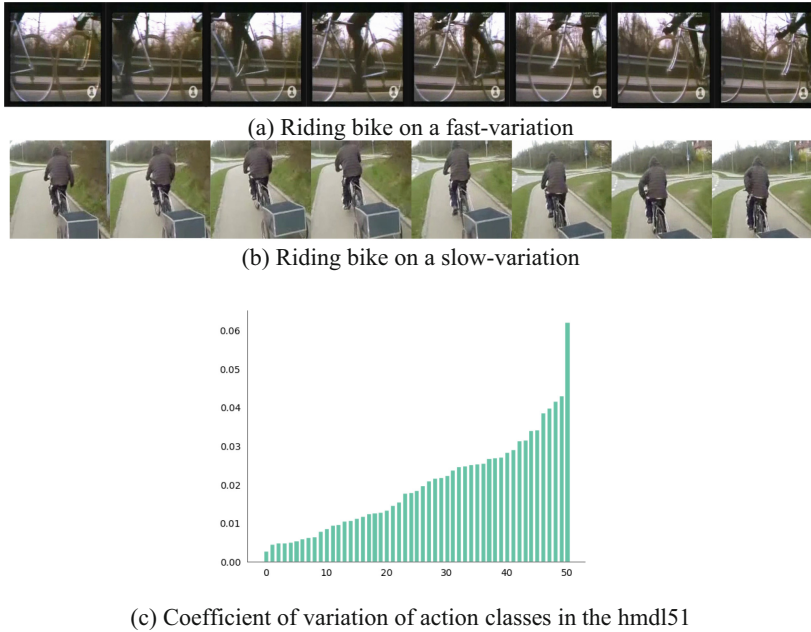


Fig. 1. Examples from the HMDB-51 dataset. The subfigures (a) and (b) show that videos tend to vary at different spatiotemporal rates for the same action (ride bike). The subfigure (c) shows the coefficients of variation of each class in the HMDB-51 dataset.

As shown in Fig. 1, the speed of riding bike in Fig. 1 (a) is faster than that in Fig. 1 (b). The action of riding a bicycle is very subtle in Fig. 1 (b), and the change on the temporal scale is very small. Note that conspicuous differences can be seen in the changes in the temporal feature. Simultaneously, the visual appearance changes in Fig. 1 (a) are also at a different rate from those in Fig. 1 (b) because of the different positions of the cameras. There are also a great number of different visual tempos variations in reality or the action recognition datasets. Figure 1 (c) show that the coefficient of variation is significant different for inter-class. For example, the fall floor has the giant view variance in the spatial-temporal frequency of the instances, while the sit-up has the smallest. We introduce the details of the coefficient of variation in Sect. 4.5 of the paper. We show that it can be exploited to improve accuracy significantly for action recognition.

Current representative models, such as $R(2 + 1)D$ [9] and GST [10], usually decompose 3D convolutions into temporal and spatial convolutions and stack them. Although, as the number of layers increases, so does their receptive field. However, adapting to different rhythms in a single model is challenging. These models struggle to cope with the identification challenges posed by various frequency variations inter-classes and intra-classes. On the other hand, redundant model parameters inevitably lead to difficulties in model training and computational burden. To extract the multi-scales feature of action instances, previous works [11–13] mainly rely on constructing a frame pyramid for the visual tempo. These methods obtain different spatial scales through splicing the feature of different layers of the backbone network. Other methods [14, 15] obtain different

temporal scales by sampling the input frames from different stride. However, they due to need to use additional models to experience variation at different visual tempos. As can be seen, the development of action recognition remains an ongoing challenge due to strict requirements for learning dynamic features of visual tempos that need a model with good perform and low cost.

In this paper, we introduce a novel and concise Multi-Receptive Field SpatioTemporal (MRF-ST) network to tackle the problem above. Similar to the decomposition convolution, we first divide the 3D convolution into temporal convolution and spatial convolution, and then implement them by two dilated convolutions [16] with different dilation rates. We realize a two-path unit Multi-Receptive Field Temporal (MRF-T) and Multi-Receptive Field Spatial (MRF-S), both of which achieve various visual tempos at the same unit. Our MRF-ST network can fuse different receptive fields for the spatiotemporal feature on the same layer without extra parts. Our major contributions can be summarized as follows:

- Firstly, we propose a new 3D convolution decomposition method, Based on our exploration of visual tempo, that can effectively model the spatial and temporal information of different receptive fields.
- Next, we can capture different visual rhythm features and model their relationships from the proposed MRF-ST network using multiple receptive fields. To the best of our knowledge, this is the first action recognition unit that simultaneously fusions different visual tempo features in the same layer of the network. In this way, dynamic characteristics can be captured more robustly.
- Then, the method utilizes an attention mechanism to assign different weights according to different contributions of different receptive fields. This allows for a more efficient adaptation to different visual tempos.
- Lastly, we evaluate MRF-ST on two action benchmarks (HMDB-51 [17] and UCF-101 [18]). Experimental results show that MRF-ST significantly improves performance. We further analyze the contribution of MRF-ST to learning visual tempos and it achieves stellar performance on several datasets with much less parameters.

The rest of this paper is organized as follows. We introduce related work progress and our advantages in Sect. 2. In Sect. 3, we detail the MRF-ST network. We perform experiments and analysis of our model in Sect. 4. The summary and outlook of this paper are in Sect. 5.

2 Related Work

In this section, we introduce the related work of action recognition in the era of deep learning networks. In particular, we discuss the work related to visual tempo in the final.

2.1 Deep Learning in Action Recognition

The related work can be divided into two categories for video action recognition. Methods in the first category often adopt a 2D + 1D paradigm, where 2D CNNs are applied

over per-frame inputs, followed by a 1D module that aggregates per-frame features. Temporal relational networks [19, 20] explored the temporal relation between learning and reasoning video frames. In particular, moving features along the temporal dimension, the method [21] only maintains the complexity of 2D CNN while achieving the performance of 3D CNN without optical flow. [22, 23] explored different fusion models for action recognition. [24, 25] studied the sequential models based on RNN and LSTM for video. For 2D CNNs deployed in these methods, the semantics of input frames cannot interact with each other early on, which limits their ability to capture visual rhythmic dynamics.

Methods [10] in the second category alternatively apply 3D CNNs that stack 3D convolutions to jointly model temporal and spatial semantics. To capture the spatiotemporal information from multiple adjacent frames, a 3D convolutional kernel [26, 27] is mainly utilized in several deep neural networks instead of a two-dimensional (2D) convolutional one. However, 3D convolution brings more parameters than 2D, making it difficult to train and requiring more hardware resources. The heavy calculation requirement and the great number of parameters are still two burdens for 3D CNN development. CoST [28] learns spatial appearance and temporal motion information using 2D convolution with weight sharing to capture three orthogonal views from video data. [29] captures spatiotemporal information from both snippet-level and long-term context by using the dilated dense blocks. [30] can obtain semantic relevance in spatial and channel dimensions through two types of attention modules. In the channel module, the attention mechanism emphasizes interdependent channel characteristics by integrating the correlation characteristics among all channel maps. And the spatial module does weighted fusion at all positions, selectively aggregating the features of each position. GST [10] employs two groups to pay attention to static and dynamic feature prompts by decomposing 3D convolution into spatial and temporal convolution in parallel.

Nevertheless, it may be not able to understand the temporal and spatial dynamics of the video. Our proposed model is also inspired by the above ideas, which can effectively utilize the most active context from a broader 3D perspective. Our model can collaboratively learn the key spatiotemporal representations of different visual rhythms by fusing two different reception fields in this paper.

2.2 Visual Tempo in Action Recognition

Understanding action semantics and temporal information is a difficult task in action recognition, especially in the variety of visual rhythms. Recently, many researchers concentrate on solving this problem [14, 31]. [11] handles multi-rate videos by randomizing the sampling rate during training. DTPN [12] also samples frames with different frames per second to construct a natural pyramidal representation for arbitrary-length input videos. In SlowFast [14], an input-level frame pyramid structure is established to encode changes in visual rhythm, which is fast and slow networks by inputting video frames sampled at different rates. The slow network tends to capture the slow rhythm while the fast network tends to capture the fast rhythm. The different rhythm features of the two networks are fusion through lateral connections. With the frame pyramid and information fusion, SlowFast can robustly capture changes in visual rhythm. S-TPNet [15] and TPN [32] take full advantage of the temporal pyramid module. They reuse the

video features and exploit various spatial scale and temporal scale pooling approaches to efficiently obtain different spatial-grained and temporal-grained features. CIDC [33] can encode the temporal sequence information of actions into the feature maps to learn the temporal association among local features in a temporal direction fashion by introducing a directional convolution unit independent of the channel.

However, this coding scheme often extracts multiple frames or multiple middle layer features, especially when we need a large pyramid scale. Note that we could deal with the concerns about visual speed in a single network. Thus, we only need to sample frames at a single rate at the input level and deal with changes in the visual tempo at the feature level using multiple receptive fields to capture different rhythms.

3 Proposed Model

In this section, we implement a baseline by the conventional 3D convolution architectures. Then we introduce the proposed MRF-ST and discuss the differences between different networks. The MRF-ST networks can be described as a spatiotemporal architecture that operates at two various receptive fields to capture visual tempos changes. Our generic architecture has MRF-T and MRF-S units (Sec. 3.2). We use the attention mechanism to capture the contributions of different receptive field modules (Sec. 3.3). Finally, complete network architecture and some discussions are given (Sec. 3.4).

3.1 3D ConvNets in Action Recognition

To verify our ideas, we implement 3D ResNet50 network as a baseline model. The video clip sampled from a 64-frame with a temporal stride of 4 as input. For a general 3D convolutional kernel with C_i input channels and C_o output channels, T, H, W are the kernel sizes along the temporal and spatial dimensions. As shown in Table 1, we decompose the 3D convolution kernel into the temporal and spatial kernel with the sizes of $w_t \in \mathbb{R}^{C_o \times C_i \times T \times 1 \times 1}$ and $w_s \in \mathbb{R}^{C_o \times C_i \times 1 \times H \times W}$. It is worth noting that the instead model learns temporal and spatial features, rather than jointly.

3.2 Multi-receptive Field Unit

There are very helpful that learning the spatiotemporal features of different visual tempos and performing good fusion for video recognition. A good strategy should preserve the spatial and temporal information to the greatest extent and capture the interaction between features of different visual tempos. Unlike the idea of constructing two input frames with two sampling rates in the SlowFast model, we use another method to use the group convolution to integrate the temporal or spatial convolution into two parts, and then use different dilation rates to achieve the effect of learning different visual tempos information.

As shown in Fig. 2, for the temporal convolution of $w_t \in \mathbb{R}^{C_{0m} \times C_{in} \times T \times 1 \times 1}$, we split it into two dilated convolutions with different dilation rates, and similar operations are also applied to the spatial convolution. We can get MRF-T and MRF-S units, with the convolution sizes of $w_t \in \mathbb{R}^{C_{0m} \times C_{im} \times T \times 1 \times 1}$ and $w_s \in \mathbb{R}^{C_{0m} \times C_{im} \times 1 \times H \times W}$. Among them,

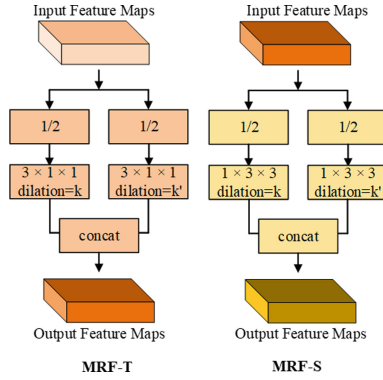


Fig. 2. Our proposed method for MFT and MFS.

Table 1. The 3D ResNet50 configuration.

Stage	Layer	Output size
conv1	$1 \times 7 \times 7, 64, \text{stride } 1 \times 2 \times 2$	$16 \times 112 \times 112$
pool1	$1 \times 3 \times 3 \text{ max, stride } 1 \times 2 \times 2$	$16 \times 112 \times 112$
res2	$\begin{bmatrix} 3 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	$16 \times 56 \times 56$
res3	$\begin{bmatrix} 3 \times 1^2, 128 \\ 1 \times 3^2, 128 \\ 1 \times 1^2, 512 \end{bmatrix} \times 4$	$16 \times 28 \times 28$
res4	$\begin{bmatrix} 3 \times 1^2, 256 \\ 1 \times 3^2, 256 \\ 1 \times 1^2, 1024 \end{bmatrix} \times 6$	$16 \times 14 \times 14$
res5	$\begin{bmatrix} 3 \times 1^2, 512 \\ 1 \times 3^2, 512 \\ 1 \times 1^2, 2048 \end{bmatrix} \times 3$	$16 \times 7 \times 7$
global average pool, FC		$1 \times 1 \times 1$

$C_{0m} = C_{0n}/2$ and $C_{im} = C_{in}/2$, in other words, m is half of n . Take MRF-T as an example. Formally, we can formulate MRF-T as:

$$X^t = (x_1^t, x_2^t, \dots, x_m^t, x_{m+1}^t, \dots, x_n^t) \tag{1}$$

$$w_t = (w_1^t, w_2^t, \dots, w_m^t) \quad (2)$$

We denote the feature maps as X^t in the input temporal convolution, where $n = T \times H \times W$. We divide X^t along the channel into X_1^t and X_2^t .

Formally, we can formulate MRF-T as:

$$\theta_1^t(X_1^t) = X_1^t \otimes \cdot w_t \quad (3)$$

$$\theta_2^t(X_2^t) = X_2^t \otimes \cdot w_t \quad (4)$$

$$y^t = \theta_1^t(X_1^t) \oplus \theta_2^t(X_2^t) \quad (5)$$

where \otimes denotes 3D convolution, w_t is temporal convolution filters shared among the two dilation rates. The $\theta_1^t(X_1^t)$ and $\theta_2^t(X_2^t)$ are the result of convolution, and then concatenate them to get y^t .

Similarly, the above formula is the operation in the temporal convolution of MRF-T, and the spatial convolution in MRF-S is also the same.

$$y^s = \theta_1^s(X_1^s) \oplus \theta_2^s(X_2^s) \quad (6)$$

MRF-T and MRF-S can encourage each group's channels to focus on the dynamic features of different rates convenient for training. MRF-TS can thus combine various tempos features naturally. Then, the w_m is the number of parameters for multi-receptive field spatiotemporal unit, we reduce the number of parameters by reducing the input channels C_{im} and output channels C_{om} .

$$w_m = (H \cdot W + T) \cdot C_{im} \cdot C_{om} \quad (7)$$

3.3 Attention for the Multi-receptive Field Unit

To make the model more suitable for learning spatiotemporal features of visual tempo changes, we design dynamic weighting for different dilation rates, which a parameter α . Since the convolution of different receptive fields contributes differently to learning, we predict α according to the network's feature map. The attention method [45] inspired us to make the attention unit. For MRF-T, we can use the formula:

$$y_a^t = \alpha_1 \theta_1^t(X_1^t) \oplus \alpha_2 \theta_2^t(X_2^t) \quad (8)$$

$\theta_1^t(X_1^t)$ and $\theta_2^t(X_2^t)$ are the result of temporal convolution with different dilation rates. α_1 and α_2 are the weights calculated by the attention module, which is shown in Fig. 3.

Specifically, we divide the input feature map into X_1^t and X_2^t along the channel, and then send them into two temporal convolutions with different dilation rates to get $\theta_1^t(X_1^t)$ and $\theta_2^t(X_2^t)$, which are concatenated to get y^t . The above process is consistent with the MRF-T described in Sect. 3.2. The difference is that we evaluated the contribution of X_1^t and X_2^t through attention operations. As shown in Fig. 3, we first use the adaptive

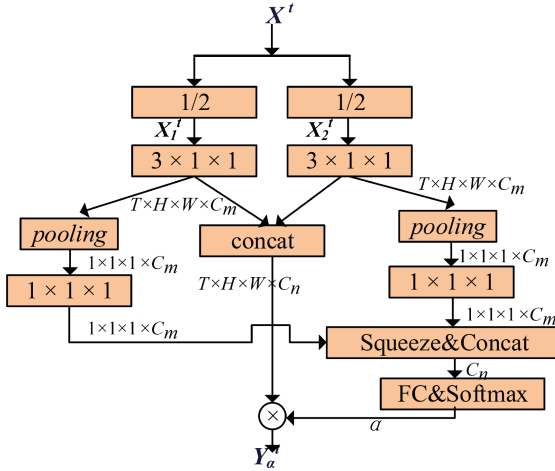


Fig. 3. Attention for the multi-receptive field unit

max pooling to reduce the feature map to $1 \times 1 \times 1 \times C_m$, along the dimensions T , H , and W . The pooled features are input into the $1 \times 1 \times 1$ convolution to capture the channels’ context information. Then, the features of two sets obtained in the previous step are concatenated and fed into a fully connected layer. This FC layer can capture the contextual information among different visual tempos. Then, we use the Softmax function to normalize the output to get α . Finally, α multiply the corresponding y^t introduced above, and the final process is formally expressed as Eq. (8). In the model, the weight coefficient of each feature depends on itself.

3.4 Network Architecture

Here we introduce the network architecture of MRF-ST for action recognition. To better study the different receptive field of visual tempo, we have proposed the following structure.

The proposed MRF-ST unit is flexible and can easily replace the convolution modules in most current networks. As shown in Fig. 4, we put the proposed units on the backbone network composed of 3D ResNet50. In our architecture, we enter 16 video clips into the network to obtain the final classification result.

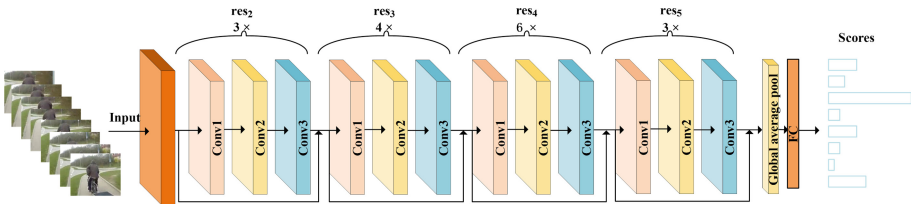


Fig. 4. The architecture of a multi-receptive field spatiotemporal network. We build it based on 3D ResNet50.

To verify the temporal and spatial effect, we design different models, as shown in Fig. 4. Compared with 3D ResNet50, we replace the temporal convolution and spatial convolution with MRF-T and MRF-S, respectively. First, we apply a separate spatial-temporal unit for baseline, which consists of three layers i.e., a 3 temporal convolution layer, a $3 \times 1 \times 1$ temporal convolution, a $1 \times 3 \times 3$ spatial convolution, and a $1 \times 1 \times 1$ convolution layer, namely Conv1, Conv2 and Conv3. Then, we just replace Conv1 with MRF-T, in order to demonstrate the effect of fusion of multiple receptive fields in the temporal dimension. Similarly, we replace Conv2 with MRF-S, to fuse multiple receptive fields in spatial dimension. Finally, we consider the combined impact of two factors and replace both Conv1 and Conv2. These are four units with comparison, and we conduct ablation study on them in Sect. 4.3.

4 Experiment

To evaluate the performance of the proposed MRF-ST network on visual tempo feature learning and action recognition tasks, we conduct experiments and ablation studies on two wild datasets, UCF-101[17] and HMDB-51[18]. In this section, we introduce the implementation details of the experiments involved in this paper. We discuss the effectiveness of each component and the correctness of the visual tempo learning concept.

4.1 Datasets

UCF-101 and HMDB-51 have been very popular in research, which are challenging benchmarks for action recognition. Table 2 lists the general information of the datasets we used.

Table 2. Details of the datasets used for evaluation. The Clips show the total number of short video clips extracted from the Videos available in the dataset..

dataset	# Clips	# Videos	# Classes
UCF-101	13320	2500	101
HMDB-51	6766	3312	51
Kinetics400	306,245	306,245	400

There are 2500 videos and 13320 clips with 101 classes from YouTube in the UCF-101. The short clips are extracted from the videos available. In this dataset, 25 people are performing each behavior, and each of them performs multiple operations. The UCF-101 dataset offers the greatest diversity in action, with large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered backgrounds, lighting conditions, and more.

The HMDB-51 with 51 different categories is mainly collected from movies. There are 3312 videos and 6766 clips, which contain a lot of facial actions and object interaction. We can divide HMDB-51 into five action categories: Common facial actions, Complex

facial movements, Common body movements, Complex body movements, and Multi-person interactive body movements.

They all have three training and testing splits for action recognition. Unless otherwise stated, our final result is the average of these splits. Our models are pre-trained on a large dataset kinetics400 [34], which contains 236763 training videos and 19095 validation videos.

4.2 Implementation Detail

In our model, we sample 64 consecutive frames from the clip and then take one every three frames to get 16 frames as input. During the inference time, we do the random crop for each frame. We use PyTorch to implement our networks, which train on the TITAN Xp GPU machine. We train the models using the CrossEntropyLoss function and the SGD optimizer. Meanwhile, we set the momentum to 0.9, the weight decay to 0.0001, and the minimum batch size to 16. We train 80 epochs to optimize all models. The learning rate is initialized to 0.01 and reduced by a factor of 10 every 30 epochs. The total training epochs are about 80.

4.3 Comparisons with the State-Of-The-Arts

Table 3. Top-1 accuracy performance on UCF-101 and HMDB-51 compared with state-of-the-art methods. All accuracies are averaged over three splits.

Method	UCF-101	HMDB-51
C3D [27]	82.3	51.6
C3D + IDT[27]	90.4	-
P3D [35]	88.6	-
STC-ResNet 101[36]	93.7	66.8
3D ResNeXt-101[37]	94.5	70.2
I3D RGB [38]	95.1	74.3
MiCT-Net [39]	88.9	63.8
MiCT-Net two-stream[39]	94.7	70.5
Two-stream ConvNet[40]	88.0	59.4
Two-Stream Fusion [41]	92.5	65.4
TSN [42]	94.9	71.0
TSM[43]	94.5	70.7
STM[44]	96.2	72.2
Our MRF-ST	96.3	73.1

We evaluate the Top-1 accuracy of our MRF-ST module embedded in the 3D ResNet-50 network against state-of-the-art methods. As Table 3 shows, our method approaches

the best performance compared with the state-of-the-art methods using only RGB, such as 3D methods [27, 38] and temporal shift methods [43, 44]. And it is also close to the results of using optical flow, such as Two-stream based methods [40, 41].

We can infer from the table that the use of optical flow information can effectively improve the competitiveness of the model. However, optical flow needs to be pre-calculated and stored on the hard disk, which requires a lot of costs. There is also not conducive to the application and migration of the model. Our model is easy to replace by the 3d convolution model to achieve a competitive effect without additional cost. Table 4 lists the Top-1 and TOP-5 accuracies. MRF-ST can achieve fine performance results in different split situations, which shows that our model has not lost its robustness due to the addition of the multi-receptive field module.

Table 4. Top-1 and TOP-5 accuracy on different splits of UCF-101 and HMDB-51

split	UCF-101		HMDB-51	
	Top-1 acc	Top-5 acc	Top-1 acc	Top-5 acc
1	96.51	99.86	73.73	93.46
2	96.63	99.57	73.59	93.86
3	96.02	99.59	72.09	91.31

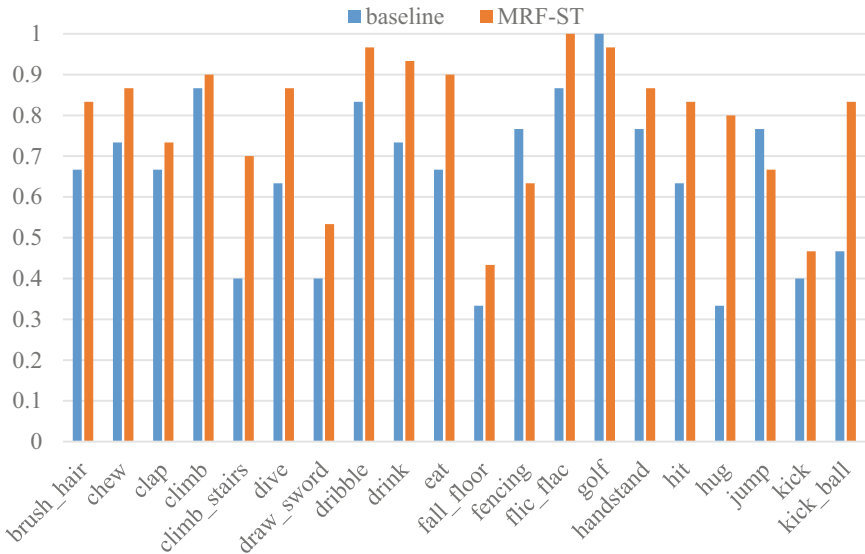


Fig. 5. Accuracy comparisons for the top-20 classes on the split1 of HMDB-51 dataset between MRF-ST (Ours) and the baseline (3D ResNet-50) model

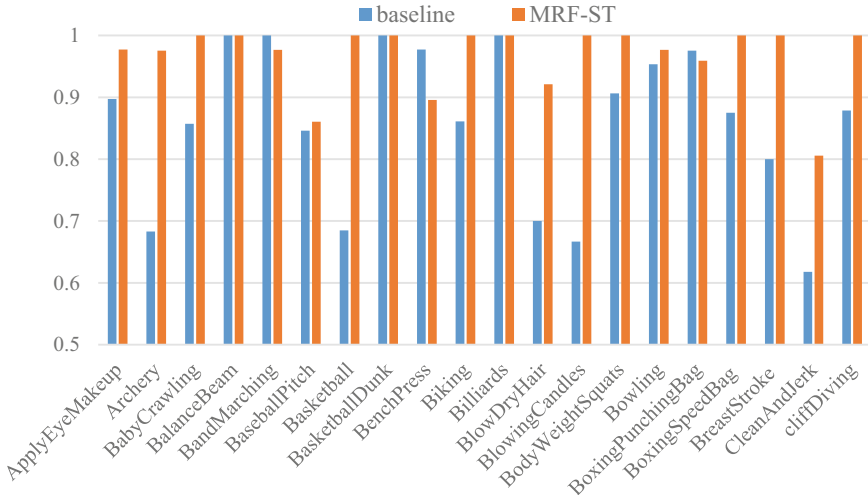


Fig. 6. Accuracy comparisons for the top-20 classes on the split1 of UCF-101 dataset between MRF-ST (Ours) and the baseline (3D ResNet-50) model

To further investigate the performance over different actions, Fig. 5 and 6 show the comparison between our model (MRF-ST) and the baseline (ResNet-50) for different categories of HMDB-51 and UCF-101 datasets. Figure 5 shows the top 20 classes’ accuracy from HMDB-51, where our model outperforms the original model. For some classes with fast visual tempo variations such as *Kick_ball* and *Dive*, our model obtains a significant performance gain (36.7% and 23.3%) over the ResNet-50 model. Our method faithfully captures the visual tempo information in both spatial and temporal dimensions. A similar conclusion can be found from Fig. 6, which demonstrates the performance improvements using our model on the UCF-101 dataset. We have achieved significant progress in most categories compared to the baseline.

4.4 Ablation Study

In this section, we conduct the ablation studies on the datasets. First, we further investigate the influence of receptive field changes on the model in Table 5. T represents the dilation rate for the temporal convolution, and S represents the dilation rate for the spatial convolution. We study that add different dilation rates to the original convolution to change the receptive field. However, we find that a single receptive field change, which, only replace with dilation convolution, cannot improve the effect or be negative. In the deep network, convolutional of different depths implicitly learn different receptive field information. At the same time, we need to perform the padding operation, when using the dilated convolution. But if the dilation rate is set too large, a lot of information will be lost. And then, we study the different units proposed in Sect. 3.2 for our action recognition model. As shown in Table 5, we improve the effect compared to the baseline by MRF-T or MRF-S units, but the effect did not increase as the dilation rate increased. Not listed in the table, when we use more dilated convolutions (more than two) to get

worse experimental results in spatial or temporal convolution. This consequence shows that the network has different receptive field fusions that can better fit the feature of visual tempos changes. However, the model will lose some feature information because the dilation rate is too large. In summary, we set the dilation rate to 1 and 2 in the MRF-T and MRF-S units.

Table 5. Top-1 accuracy performance on UCF-101 and HMDB-51 split1 compared with different dilation rates.

Model	Dilation rate	UCF-101	HMDB-51
		Top-1 acc	Top-1 acc
baseline	T = 1,S = 1	88.9	66.5
baseline	T = 2,S = 1	88.3	65.9
baseline	T = 3,S = 1	87.5	63.9
baseline	T = 1,S = 2	88.8	64.5
baseline	T = 1,S = 3	88.6	64.3
MRF-T	T = (1,2),S = 1	92.5	70.5
MRF-T	T = (2,3),S = 1	91.6	68.9
MRF-S	T = 1,S = (1,2)	91.8	68.6
MRF-S	T = 1,S = (2,3)	90.3	67.4

We show the parameters and accuracy of different models in Table 6. Their structure is shown in Fig. 4 above, but the attention module is implied. Our model has improved compared with baseline, and the parameters have been significantly reduced. For example, comparing MRF-ST with baseline, we reduce the number of parameters by 12.2×10^6 . Meanwhile, we improve the accuracy of top-1 and top-5 by 7.6% and 2.86% in UCF-101. And with more attractive results in HMDB-51, we improve by 7.19% and 5.36%. This result fully shows that our model has better results with fewer parameters.

Table 6. The parameters and accuracy of different models are compared on split1 of UCF-101 and HMDB-51.

Model	Param $\times 10^6$	UCF-101		HMDB-51	
		Top-1 acc	Top-5 acc	Top-1 acc	Top-5 acc
baseline	32.4	88.91	97.00	66.54	88.10
MRF-S	26.7	93.16	99.18	70.26	91.57
MRF-T	23.7	94.13	99.49	71.37	92.16
MRF-ST	20.2	96.51	99.86	73.73	93.46

4.5 Empirical Analysis

To verify whether MRF-ST has captured the variance of visual tempos, we used the parameter α to conduct some experimental analysis. We measure the coefficient of variation of the action instance to distinguish the visual tempos of the action instance accurately. Specifically, we calculate the cosine acquaintance of adjacent frames of the video to characterize the difference in video motion pixel level. Then we use the coefficient of variation to measure the difference in cosine similarity changes inter-classes and intra-classes. The coefficient of variation can well reflect the visual tempo. However, the model-based method, which measures the probability change of action category to express the visual tempo, will be greatly affected by its measure model of bias. Accuracy comparisons for the top-20 classes of variation coefficients, as shown in Fig. 7. Comparing with the results shown in Fig. 5, the MRF-ST model can achieve better results in categories with fast changes, which shows that our model can effectively model visual tempo.

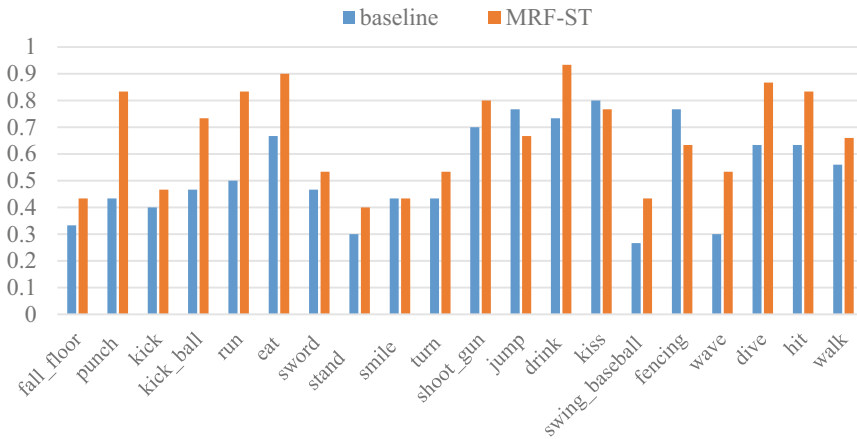


Fig. 7. Accuracy comparisons for the top-20 classes of variation coefficients on the split1 of HMDB-51 dataset between MRF-ST (Ours) and the baseline (3D ResNet50) model.

5 Conclusion

Feature learning from visual tempo variation is the unneglectable challenge in video action recognition. We propose a novel spatiotemporal feature learning operation, which learns visual tempos fusion from multiple receptive fields. Although we do not have a deeper model to achieve the best results, we verified learning visual tempos is essential. Experiments on the datasets illustrate the availability of the proposed architecture and the easiness of learning visual tempo variation features for action recognition. We hope that these explorations will inspire more video recognition to design models.

References

1. Kang, S., Wu, H., Yang, X., et al.: Discrete-time predictive sliding mode control for a constrained parallel micropositioning piezostage. *IEEE Trans. Systems, Man, Cybernetics: Systems* **52**(5), 3025–3036 (2021)
2. Zheng, Q., Zhu, J., Tang, H., et al.: Generalized label enhancement with sample correlations. *IEEE Trans. Knowledge Data Eng.* (2021)
3. Lu, H., Zhang, Y., Li, Y., et al.: User-oriented virtual mobile network resource management for vehicle communications. *IEEE Trans. Intell. Transp. Syst.* **22**(6), 3521–3532 (2020)
4. Lu, H., Zhang, M., Xu, X., et al.: Deep fuzzy hashing network for efficient image retrieval. *IEEE Trans. Fuzzy Syst.* **29**(1), 166–176 (2020)
5. Jin, X., Sun, W., Jin, Z.: A discriminative deep association learning for facial expression recognition. *Int. J. Mach. Learn. Cybern.* **11**(4), 779–793 (2020)
6. Zhuang, D., Jiang, M., Kong, J., et al.: Spatiotemporal attention enhanced features fusion network for action recognition. *Int. J. Mach. Learn. Cybern.* **12**(3), 823–841 (2021)
7. Ziaeefard, M., Bergevin, R.: Semantic human activity recognition: a literature review. *Pattern Recogn.* **48**(8), 2329–2345 (2015)
8. Chen, L., Song, Z., Lu, J., et al.: Learning principal orientations and residual descriptor for action recognition. *Pattern Recogn.* **86**, 14–26 (2019)
9. Tran, D., Wang, H., Torresani, L., et al.: A closer look at spatiotemporal convolutions for action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459 (2018)
10. Luo, C., Yuille, A.L.: Grouped spatial-temporal aggregation for efficient action recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5512–5521 (2019)
11. Zhu, Y., Newsam, S.: Random temporal skipping for multirate video analysis. In: *Asian Conference on Computer Vision*. Springer, Cham, pp. 542–557 (2018). https://doi.org/10.1007/978-3-030-20893-6_34
12. Zhang, D., Dai, X., Wang, Y.F.: Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection. In: *Asian Conference on Computer Vision*. Springer, Cham, pp. 712–728 (2018). https://doi.org/10.1007/978-3-030-20870-7_44
13. Du, Y., Yuan, C., Li, B., Zhao, L., Li, Y., Hu, W.: Interaction-aware spatio-temporal pyramid attention networks for action classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *ECCV 2018*. LNCS, vol. 11220, pp. 388–404. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01270-0_23
14. Feichtenhofer, C., Fan, H., Malik, J., et al.: Slowfast networks for video recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211 (2019)
15. Zheng, Z., An, G., Wu, D., et al.: Spatial-temporal pyramid based convolutional neural network for action recognition. *Neurocomputing* **358**, 446–455 (2019)
16. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015)
17. Kuehne, H., Jhuang, H., Garrote, E., et al.: HMDB: a large video database for human motion recognition. In: *2011 International Conference on Computer Vision*. IEEE, pp. 2556–2563 (2011)
18. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012)
19. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *ECCV 2018*. LNCS, vol. 11205, pp. 831–846. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_49

20. Shi, Y., Tian, Y., Huang, T., et al.: Temporal attentive network for action recognition. In: 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp. 1–6 (2018)
21. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7083–7093 (2019)
22. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941 (2016)
23. Zolfaghari, M., Singh, K., Brox, T.: ECO: efficient convolutional network for online video understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 713–730. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_43
24. Du, W., Wang, Y., Qiao, Y.: Recurrent spatial-temporal attention network for action recognition in videos. *IEEE Trans. Image Process.* **27**(3), 1347–1360 (2017)
25. Li, C., Zhang, B., Chen, C., et al.: Deep manifold structure transfer for action recognition. *IEEE Trans. Image Process.* **28**(9), 4646–4658 (2019)
26. Li, J., Liu, X., Zhang, M., et al.: Spatio-temporal deformable 3d convnets with attention for action recognition. *Pattern Recogn.* **98**, 107037 (2020)
27. Tran, D., Bourdev, L., Fergus, R., et al.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
28. Li, C., Zhong, Q., Xie, D., et al. Collaborative spatiotemporal feature learning for video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7872–7881 (2019)
29. Xu, B., Ye, H., Zheng, Y., et al.: Dense dilated network for video action recognition. *IEEE Trans. Image Process.* **28**(10), 4941–4953 (2019)
30. Fu, J., Liu, J., Jiang, J., et al.: Scene segmentation with dual relation-aware attention network. *IEEE Trans. Neural Networks Learning Syst.* **32**(6), 2547–2560 (2020)
31. Wang, Z., Chen, K., Zhang, M., et al.: Multi-scale aggregation network for temporal action proposals. *Pattern Recogn. Lett.* **122**, 60–65 (2019)
32. Yang, C., Xu, Y., Shi, J., et al.: Temporal pyramid network for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 591–600 (2020)
33. Li, X., Shuai, B., Tighe, J.: Directional temporal modeling for action recognition. In: European Conference on Computer Vision. Springer, Cham, pp. 275–291 (2020). https://doi.org/10.1007/978-3-030-58539-6_17
34. Kay, W., Carreira, J., Simonyan, K., et al.: The kinetics human action video dataset. arXiv preprint [arXiv:1705.06950](https://arxiv.org/abs/1705.06950) (2017)
35. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: proceedings of the IEEE International Conference on Computer Vision, pp. 5533–5541 (2017)
36. Diba, A., et al.: Spatio-temporal channel correlation networks for action classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 299–315. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_18
37. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4768–4777 (2017)
38. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)

39. Zhou, Y., Sun, X., Zha, Z.J., et al.: Mict: Mixed 3d/2d convolutional tube for human action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 449–458 (2018)
40. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Processing Syst.* **27** (2014)
41. Christoph, R., Pinz, F.A.: Spatiotemporal residual networks for video action recognition. *Adv. Neural Inf. Processing Syst.* 3468–3476 (2016)
42. Wang, L., Xiong, Y., Wang, Z., et al.: Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(11), 2740–2755 (2018)
43. Lin, J., Gan, C., Han, S.: Temporal shift module for efficient video understanding. *CoRR* abs/1811.08383 (2018)
44. Jiang, B., Wang, M.M., Gan, W., et al.: Stm: Spatiotemporal and motion encoding for action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2000–2009 (2019)
45. Yin, M., Yao, Z., Cao, Y., et al.: Disentangled non-local neural networks. In: European Conference on Computer Vision. Springer, Cham, pp. 191–207 (2020). https://doi.org/10.1007/978-3-030-58555-6_12
46. Huimin, L., Zhang, M., Xu, X.: Deep fuzzy hashing network for efficient image retrieval. *IEEE Trans. Fuzzy Syst.* **29**(1), 166176 (2020). <https://doi.org/10.1109/TFUZZ.2020.2984991>
47. Huimin, L., Li, Y., Chen, M., et al.: Brain intelligence: go beyond artificial intelligence. *Mobile Networks Appl.* **23**, 368–375 (2018)
48. Huimin, L., Li, Y., Shenglin, M., et al.: Motor anomaly detection for unmanned aerial vehicles using reinforcement learning. *IEEE Internet Things J.* **5**(4), 2315–2322 (2018)
49. Huimin, L., Qin, M., Zhang, F., et al.: RSCNN: A CNN-based method to enhance low-light remote-sensing images. *Remote Sensing* **13**(1), 62 (2020)
50. Huimin, L., Zhang, Y., Li, Y., et al.: User-oriented virtual mobile network resource management for vehicle communications. *IEEE Trans. Intell. Transp. Syst.* **22**(6), 3521–3532 (2021)