High Performance Gesture Recognition via Effective and Efficient Temporal Modeling

Yang Yi[1*†], **Feng Ni**[2*], **Yuexin Ma**[3], **Xinge Zhu**[4], **Yuankai Qi**[5], **Riming Qiu**[1], **Shijie Zhao**[1], **Feng Li**[1] and **Yongtao Wang**[2]
[1] Media Lab, Tencent
[2]Peking University
[3]University of Hong Kong
[4]The Chinese University of Hong Kong
[5]Harbin Institute of Technology, Weihai, China

Abstract

State-of-the-art hand gesture recognition methods have investigated the spatiotemporal features based on 3D convolutional neural networks (3DCNNs) or convolutional long short-term memory (ConvLSTM). However, they often suffer from the inefficiency due to the high computational complexity of their network structures. In this paper, we focus instead on the 1D convolutional neural networks and propose a simple and efficient architectural unit, Multi-Kernel Temporal Block (MKTB), that models the multi-scale temporal responses by explicitly applying different temporal kernels. Then, we present a Global Refinement Block (GRB), which is an attention module for shaping the global temporal features based on the cross-channel similarity. By incorporating the MKTB and GRB, our architecture can effectively explore the spatiotemporal features within tolerable computational cost. Extensive experiments conducted on public datasets demonstrate that our proposed model achieves the state-of-the-art with higher efficiency. Moreover, the proposed MKTB and GRB are plug-and-play modules and the experiments on other tasks, like video understanding and video-based person re-identification, also display their good performance in efficiency and capability of generalization.

1 Introduction

Gesture recognition is a longstanding topic in computer vision, whose goal is to assign the corresponding label to hand gesture video. It plays an important role in many real applications, such as video surveillance, human-computer interaction, etc. A variety of methods [Narayana *et al.*, 2018; Zhang *et al.*, 2018; Zhang *et al.*, 2017; Miao *et al.*, 2017] have been developed in recent years.

Considering the nature of sequential data, many gesture recognition approaches focus on extracting discriminative spatiotemporal features, including 3D convolutional neural networks (3DCNNs) [Miao *et al.*, 2017], the ensemable of 2D convolutional neural networks (2DCNNs) [Narayana *et al.*, 2018] and convolutional long short-term memory (ConvLSTM) [Zhang *et al.*, 2017]. However, the huge computational cost in previous methods can cause severe inefficiency in real-world deployment. In this work, we make use of the 1D convolutional neural networks (1DCNNs used in both MKTB and GRB) for exploring temporal information as it is much more efficient and applicable. In particular, we apply multiple 1DCNNs with different kernel sizes to extract pyramidal temporal features. As Figure. 1 shows, our proposed model is better and faster than current state-of-the-art methods that utilize 3DCNNs and ConvLSTM.

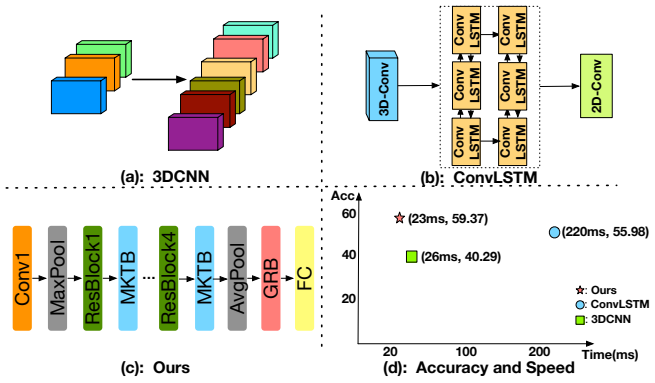

Figure 1: We display the pipeline of 3DCNN, ConvLSTM and our method in (a), (b) and (c), respectively. (d) shows the comparison with other frameworks on accuracy and speed, and it can be found that our model outperforms 3DCNNs or ConvLSTM methods.

In this paper, we first introduce a simple and effective architecture unit, *i.e.*, Multi-Kernel Temporal Block (MKTB), to model the multi-scale temporal information. It consists of multiple 1D convolutional kernels with different sizes in a depthwise fashion, thus enjoying high efficiency. Experiments further demonstrate that MKTBs bring significant improvements in accuracy. Furthermore, we design a Global Refinement Block (GRB) to adaptively explore the high-level temporal features by modeling the cross-channel similarity,

[*] indicates equal contributions.
[†]indicates corresponding author:lisiyi@tencent.com.

which is neglected in the MKTB. It performs as an attention mechanism and allows distant temporal features to contribute to the filtered temporal response at a location based on cross-channel similarity. Therefore, there are several advantages of our proposed modules: (1) In contrast to the complex recurrent and 3D convolutional operations, MKTB captures both short-term and long-term temporal information by using the multiple 1D depthwise convolutions. As shown in experiments, they are efficient; (2) The proposed modules, MKTB and GRB, maintain the same size between input and output, and can be easily deployed everywhere. Overall, the cooperation of these two components leads to a temporal modeling process that holds the discriminative spatiotemporal information effectively and efficiently.

We test the proposed model on various public datasets, including IsoGD [Wan *et al.*, 2016] and Jester [TwentyBn, 2017]. On these experiments, the proposed model yields considerable improvement over existing methods, about 3.39% in accuracy and $10\times$ in efficiency. Extensive ablation studies are also conducted to show the effectiveness of different components. Because the proposed MKTB and GRB are plug-and-play modules and gesture recognition mainly focuses on hand region, we extend these modules to video understanding task which models human-object relationships and video-based re-identification task, obtaining notable performance on Something-Something-V1 dataset [Goyal *et al.*, 2017] and MARS [Zheng *et al.*, 2016], respectively, which demonstrates its strong generalization capability and scalability.

2 Related Work

2.1 Temporal Modeling for Action Recognition

Recently, many temporal modeling approaches for action recognition achieve remarkable success. Based on 2DCNN, Temporal Segment Network (TSN) [Wang *et al.*, 2016] models long-range temporal structures with segment-based sampling and aggregation module. To learn spatial and temporal information jointly, 3DCNN [Miao *et al.*, 2017] and its variants have been widely adopted. C3D [Li *et al.*, 2016] designs a 3DCNN with small $3 \times 3 \times 3$ convolution kernels to learn spatiotemporal features, while I3D [Carreira and Zisserman, 2017] inflates the convolutional filters and pooling kernels into 3D structures. Considering the tradeoffs between effectiveness and efficiency, P3D [Qiu *et al.*, 2017] and R(2+1)D [Tran *et al.*, 2018] decompose 3D convolutions into separate spatial and temporal convolutions. In addition, non-local network [Wang *et al.*, 2018] presents non-local operations to capture long-range dependencies. However, these 3DCNN based methods suffer from the huge computational cost, often causing the inefficiency.

Differently, gesture recognition mainly focuses on the hand region instead of human-human or human-object relationships and temporal information matters in this task.

2.2 Gesture Recognition

For gesture recognition, deep learning based approaches have become commonplace nowadays. According to the basic unit of model, these methods can be grouped into the following three categories.

It is straightforward to apply 2DCNN as feature extractor for gesture recognition. [Narayana *et al.*, 2018] fuses multi-channels(*i.e.*, global/left-hand/right-hand for RGB/depth/RGB-flow/depth-flow modalities) and utilizes separate 2DCNN for each channel with stacked frame sequences as input. With an ensemble of multiple weak models, this method obtains impressive accuracy.

To take into account the temporal information, most of existing models resort to 3DCNNs. [Zhu *et al.*, 2016; Li *et al.*, 2016; Miao *et al.*, 2017] employ different 3DCNN architectures (*e.g.*, C3D and ResC3D) to learn spatiotemporal features.

[Zhang *et al.*, 2017] combines 3DCNN, bidirectional ConvLSTM and 2DCNN into a unified framework, in which the 3DCNN and ConvLSTM focus on short-term and long-term spatiotemporal information respectively. [Zhang *et al.*, 2018; Zhu *et al.*, 2017; Ma *et al.*, 2018] propose several variants of LSTM to explore the attention mechanism in ConvLSTM.

Typical approaches that utilize 3DCNN or ConvLSTM suffer from high computational complexity and memory consumption. In this paper, we exploit critical temporal information for gesture recognition task by introducing efficient and plug-and-play modules based on existing 2D architectures.

3 Methodology

To exploit critical temporal information and maintain high efficiency, our strategy is to insert modules that learn spatiotemporal features into existing models. In this section, we first present the full pipeline of our approach. Then we introduce the proposed modules, *i.e.*, Multi-Kernel Temporal Block (MKTB) and Global Refinement Block (GRB), and show how we integrate these modules into existing 2DCNN.

3.1 Framework Overview

Considering the efficiency and flexibility, our model builds upon the popular TSN framework [Wang *et al.*, 2016] , in which sparsely sampled frames are passed through a 2D backbone network followed by a consensus (aggregation) function (*e.g.*, MaxPooling). Although TSN models long-range temporal structures with the segment-based sampling and aggregation modules from input space, it lacks of capability of modeling the temporal information from feature-space. Since the proposed MKTB and GRB are effective temporal modeling modules in feature-space, they are complementary to TSN.

The overall pipeline of the proposed method is depicted in Figure. 1(c). Inherited from TSN, the input video is divided into T temporal segments. One frame is randomly sampled from each segment, so the size of input tensor is $(B \times T) \times 3 \times H \times W$, where B, H and W is the batch size, height and width of input, respectively. In this paper, T is set to 8. It can be found that our pipeline is the combination of new modules and classic residual network, which only incurs minor additional computation as the new modules are computationally efficient.

3.2 Multi-Kernel Temporal Block

To achieve the effective and efficient temporal modeling, the proposed MKTB has two appealing properties: (1) Unlike the

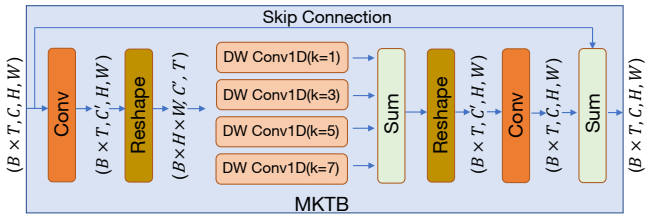

Figure 2: The details of proposed MKTB. Depthwise 1D convolution (DW Conv1D) is used in MKTB.

existing methods that employ 3DCNNs, in which they often perform convolutional operation for both spatial and temporal dimension jointly, our proposed MTKB decouples the joint spatial-temporal modeling process and focuses on learning the temporal information; (2) The design of multi-kernel also works well on shaping the pyramidal and discriminative temporal features, which significantly boosts the performance.

We denote the feature maps from layer l of 2DCNN (ResNet-50 is used in this paper) as $F_s \in R^{(B\times T)\times C\times H\times W}$. To save the computational cost in multi-kernel temporal convolutions, we reduce the channels of F_s via a convolution layer with kernel size of 1×1, obtaining a new feature representation, denoted as $F_s^{'} \in R^{(B\times T)\times C^{'}\times H\times W}$. Then $F_s^{'}$ is reshaped to $F_t^{'} \in R^{(B\times H\times W)\times C^{'}\times T}$ before the temporal convolution. Formally, the 1D convolution is defined as:

$$Y = b + \sum_{0}^{C^{'}-1} w * F_t^{'}, \tag{1}$$

where $*$ denotes the multiplication operator, $C^{'}$ is the number of input channels, b is the bias term and w denotes the weights of conv1D. As shown in Figure. 2, to further preserve efficiency in MKTB we utilize depthwise [Chollet, 2017] temporal convolution to perform computation independently over each channel, which means that the MKTB block focuses on modeling temporal information for each channel. The pyramidal features are fused by element-wise summation, followed by a reshaping operation. Another 1×1 convolution is connected at the end to keep the output having the same number of channels as input. Besides, a skip connection is added to facilitate model training.

3.3 Global Refinement Block

MKTB mines the pyramidal temporal features with depthwise 1D convolution over each channel separately, which mainly focuses on the local neighborhoods (the size is related to the kernel size). However, the global temporal features across channels are not sufficiently attended. Inspired by the non-local mechanism [Wang *et al.*, 2018], which computes the response at a position as a weighted sum of features at all positions, the GRB is designed to perform the weighted temporal aggregation, in which it allows distant temporal features to contribute to the filtered temporal features according to the cross-channel similarity.

The pipeline of GRB is shown in Figure. 3. The red dash square represents the cross-channel similarity based on the

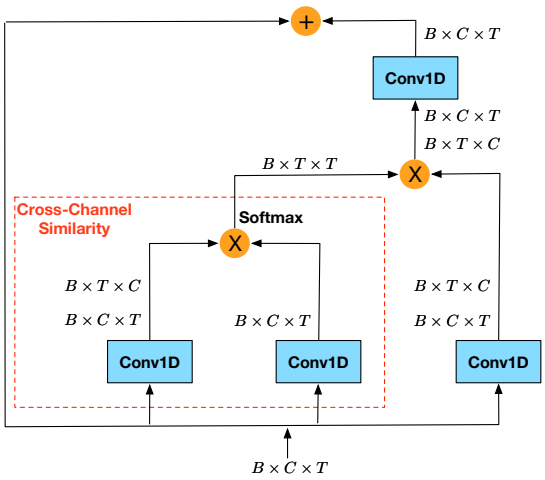

Figure 3: The pipeline of Global Refinement Block. Red dash square represents the cross-channel similarity modeling.

temporal features. Although the workflow of GRB is inherited from non-local network, it is worthy to stress some differences between the non-local module and GRB: (1) Unlike the non-local network which performs the filtering on spatial domain, GRB focuses on the temporal attention modeling, and it works on the last stage rather than the earlier part (earlier part for non-local network); (2) The 2D convolution in non-local module is replaced with 1D temporal convolution in GRB, which naturally suits the temporal features with a more efficient manner. Besides, we keep the number of channels for each convolution layer in GRB due to the lightweight input, which decouples the spatial domain.

4 Experiments

4.1 DataSets

IsoGD. IsoGD [Wan *et al.*, 2016] is a large-scale multi-modality gesture dataset which contains 249 gesture classes. In total, there are 47,933 gesture videos for each modality. This database is split into three sub-datasets: 35,878 videos for training, 5,784 videos for validation and 6,271 videos for testing.

Jester. Jester[TwentyBn, 2017] is a large collection of densely-labeled video clips of hand gestures, containing 148,092 gesture videos performed by workers in front of a laptop camera or webcam. There are 27 gesture classes, each of which has more than 5,000 instances on average, making this dataset more indispensable for gesture recognition.

Something-Somthing-V1. Something-Somthing-V1[Goyal *et al.*, 2017] is a challenging dataset that shows basic actions with everyday objects. Temporal information also plays a key role in this dataset, so we conduct extensive experiments to verify the effectiveness and generalization of the temporal modeling modules.

MARS. MARS[Zheng *et al.*, 2016] is the largest video-based person re-identification dataset. To further evaluate our method, we perform experiments on this task where temporal

information helps to perform more accurate pedestrian alignments.

4.2 Implementation details

Training. In this paper, we use ResNet-50 [He *et al.*, 2016] pre-trained on ImageNet [Deng *et al.*, 2009] as the 2D backbone. Unless otherwise noted, we set temporal segments $T = 8$. Following data augmentation strategies of TSN [Wang *et al.*, 2016], the frames are cropped and resized to 224×224 after aspect ratio jittering and scale jittering. For all experiments, we adopt mini-batch SGD to optimize the model with momentum of 0.9 and weight decay of $5e^{-4}$. We train for 60 epochs with cross entropy loss and batch size of 48. The learning rate is initialized as 0.01 and reduced by a factor of 10 every 20 epochs. Dropout layer with ratio of 0.5 is added before the classification layer. The proposed networks are trained with PyTorch deep learning framework on GPUs of NVidia Tesla P40 with CUDA 8.0. The code will be available at https://github.com/nemonameless/Gesture-Recognition.

Inference. During inference, we take the efficiency into consideration by uniformly sampling the same number of frames as in training stage. Simple center crop and scale operations are used during preprocessing.

4.3 Comparison with the State-of-the-Art

Table 1 shows the comparison with the state-of-the-art results on IsoGD dataset.

First, we compare the performance of our proposed model with method [Narayana *et al.*, 2018](referred to as 2D-ResNet50). For fair comparison, we report the results of global channel in [Narayana *et al.*, 2018]. Both methods adopt 2DCNN as backbone network. It can be seen that our model outperforms the 2D-ResNet50 by significant margin. The 2D-ResNet50 model simply takes the stacked images as input, lacking of temporal modeling structures. In contrast, the proposed MKTB and GRB modules learn effective spatiotemporal feature maps at different stages of network.

Then, compared with 3DCNN models, including [Li *et al.*, 2016](referred to as C3D), [Zhu *et al.*, 2016](referred to as Pyramidal-C3D) and [Miao *et al.*, 2017](referred to as ResC3D), the proposed model achieves more than 10% performance gain over these 3DCNNs on each modality, which demonstrates that the spatiotemporal representation learned by MKTB and GRB modules is more effective than ordinary 3DCNNs. More specifically, instead of learning spatiotemporal information simultaneously as in 3DCNNs, we ease the learning process by using 2DCNN and 1D temporal convolution to learn spatial and temporal information separately. In terms of efficiency, 3DCNNs typically suffers from high memory and computation cost. Our network utilizes 2D convolution and 1D temporal convolution to reason spatial and temporal structures while maintaining complexity close to the backbone network. Besides, compared with the recent methods [Zhang *et al.*, 2017](referred to as R3D-BiCLSTM-2D) and [Zhang *et al.*, 2018](referred to as R3D-AttCLSTM-2D) that utilize cascaded 3DCNN, ConvLSTM and 2DCNN to capture short-term and long-term spatiotemporal information, the proposed method consistently outperforms them by 8.06% and 3.39% on RGB modality, respectively. Similar scenarios can also be observed on depth and flow modality.

Model	RGB	Depth	Flow	Fusion
2D-ResNet50	33.22	27.98	46.22	61.40
Pyramidal C3D	36.58	38.00	-	45.02
C3D	37.30	40.50	-	49.20
ResC3D	45.07	48.44	44.45	64.40
R3D-BiCLSTM-2D	51.31	49.81	45.30	58.65
R3D-AttCLSTM-2D	55.98	53.28	46.51	-
Ours	**59.37**	**58.97**	**58.90**	**72.11**

Table 1: Comparison with the state-of-the-art results on validation set of IsoGD dataset.

Model	Top1 Acc. (%)
TSN	81.45
MFFs(8-MFFs-0f1c)	92.90
2-frame TRN	75.65
5-frame TRN	91.40
Multiscale TRN	93.70
R3D-AttCLSTM-2D (variant-a)	95.08
R3D-AttCLSTM-2D (variant-c)	95.13
TSM	94.40
Ours	**95.15**

Table 2: Results on the validation set of Jester dataset.

As for the multimodal evaluation, we simply fuse the predicted probabilities on each modality for our method. From the last column we can see that our method achieves better performance in multimodal settings.

Table 2 compares the proposed network to other recent methods on validation set of Jester dataset. Methods list in the table use only RGB modality without extra information such as optical flow. As can be seen, the proposed method achieves promising performance compared to the state-of-the-art methods. In particular, our model obtains higher accuracy than TSM [Lin *et al.*, 2018] (single crop evaluation is performed) which is the recent state-of-the-art method in the field of video understanding. Besides, our model consistently outperforms the TRN network [Zhou *et al.*, 2018] and MFFs

Model	Input	Speed(VPS)	Top-1
I3D	$8 \times 3 \times 224 \times 224$	39.37	36.71
P3D-C	$8 \times 3 \times 224 \times 224$	37.31	40.29
C3D	$32 \times 3 \times 112 \times 112$	11.11	37.30
TSN	$8 \times 3 \times 224 \times 224$	54.34	36.15
AttConvLSTM	$32 \times 3 \times 112 \times 112$	4.50	55.98
Ours	$8 \times 3 \times 224 \times 224$	43.47	59.37

Table 3: Speed-Accuracy comparison on validation set of IsoGD. For speed comparison, we implement the first three models with PyTorch . AttConvLSTM model is provided by the authors with TensorFlow implementation.

Model	Backbone	Frames	Top-1
TSN	BNInception	8	19.5
TRN-Multiscale	BNInception	8	34.4
I3D	3D-ResNet-50	64	41.6
ECO(Kinetics)	BNInception+3D-ResNet-18	8	39.6
ECO(Kinetics)	BNInception+3D-ResNet-18	16	41.4
TSM(Kinetics)	ResNet-50	8	43.4
Ours	ResNet-50	8	42.1

Table 4: Performance comparison on Something-Something-V1 dataset. Only RGB modality is evaluated for comparison. 'Kinetics' indicates a model pre-trained on Kinetics action dataset instead of ImageNet dataset.

Model	#Frame	MAP	Rank-1	Rank-5
TSN	8	58.4%	69.9%	85.7%
Ours	8	63.5%	74.2%	88.3%

Table 5: Performance comparison on the testing set of MARS.

network [Kopuklu *et al.*, 2018] with a notable margin. Our model achieves similar accuracy compared to [Zhang *et al.*, 2018] but about $10\times$ in efficiency as shown in Table 3.

The proposed model not only achieves promising performance on different datasets but also maintains competitive efficiency. We show the VPS(videos per second) and accuracy in Table 3. Notably, our model is around $10\times$ faster than recent state-of-the-art [Zhang *et al.*, 2018] with higher accuracy (3.39% performance gain). Compared to deeper 3DCNNs, including I3D [Carreira and Zisserman, 2017] and P3D [Qiu *et al.*, 2017], our model achieves significant improvements in accuracy and delivers faster processing speed. Although our model shows lower efficiency than TSN method, it achieves huge advantage from a accuracy perspective. Therefore, our model achieves better trade-off towards efficiency and accuracy.

4.4 Extensive experiments beyond gesture recognition

Since gesture recognition mainly focuses on the hand region, we conduct extensive experiments on other video tasks to verify the effectiveness and generality of our temporal modeling modules.

Something-Something-V1. Different from gesture recognition, Something-Something-V1 is a video understanding dataset which models complex human-object relationships where temporal reasoning plays an important role. The model in this experiment is the same with that used for gesture recognition. Table 4 shows that our model is very competitive compared to recent methods TSN[Wang *et al.*, 2016], TRN [Zhou *et al.*, 2018], ECO [Zolfaghari *et al.*, 2018], I3D [Carreira and Zisserman, 2017] and TSM [Lin *et al.*, 2018]. In particular, the proposed model achieves 22.6% gain than TSN and 7.7% than MultiScale TRN. Given 8 frames as input, our model outperforms ECO model by 2.5%. Notably, our model pre-trained on ImageNet is slightly worse than TSM which is pre-trained on Kinetics [Carreira and Zisserman, 2017], a large-scale action classification dataset.

Mothed	Position	Top-1 Acc. (%)	Speed
TSN	-	36.15	54.34
TSN+MKTB	$res2$	45.60	48.03
TSN+MKTB	$res3$	46.88	51.89
TSN+MKTB	$res4$	46.16	52.56
TSN+MKTB	$res5$	40.08	52.63
TSN+MKTB	$res4, res5$	46.47	50.79
TSN+MKTB	$res3, res4$	48.91	49.50
TSN+MKTB	$res2, res4, res5$	52.25	45.15
TSN+MKTB	$res3, res4, res5$	50.50	47.81
TSN+MKTB	$res2, res3, res4, res5$	53.10	43.47

Table 6: Comparison between different positions for MKTB modules. 'Position' indicates the place right after residual block.

Video-based person re-identification. For video-based person re-identification task, effective temporal modeling facilitates the pedestrian alignments, thus producing accurate spatiotemporal representation for video sequences. To evaluate our method, we perform comparative experiments on the largest video-based person re-identification dataset MARS [Zheng *et al.*, 2016]. Specifically, we consider our model and TSN as feature extractors and we utilize simple cross entropy loss on the training set of MARS. During testing, we simply sample 8 frames and compute the euclidean distance of spatio-temporal representations between query and gallery. As shown in Table 5, our method boosts the accuracy of video-based person re-identification by incorporating the efficient and effective temporal modeling modules.

4.5 Ablation Study

In this section, we perform extensive ablation studies to give more insight of our proposed model on IsoGD dataset.

Where to add MKTB. Table 6 shows the impact of the different positions for MKTB modules in whole pipeline. Since feature maps from different layers represent different patterns(*e.g.*, high-level feature maps typically represent semantic information), adding more MKTB modules facilitates the network to deal with scale variation of spatiotemporal features. Placing MKTB modules right after $(res2, res3, res4, res5)$ shows the best performance, demonstrating the importance of modeling temporal information. We can see that MKTB module after $res2, res3$, or $res4$ alone brings similar gains while MKTB after $res5$ delivers smallest gains. Besides, $(res2, res4, res5)$ is better than $(res3, res4, res5)$. We conjecture that feature maps in lower level provide more spatial information, which facilitates the attention converging to hand region. It is worth pointing out that we reduce the number of channels before sending into the MKTB module, suggesting that the computation cost spent on different positions is similar.

Instantiations of MKTB and GRB. We propose several variants of MKTB which still maintain the multi-scale property, but in different ways, to investigate the effect of this property. These variants utilize different dilation rates to

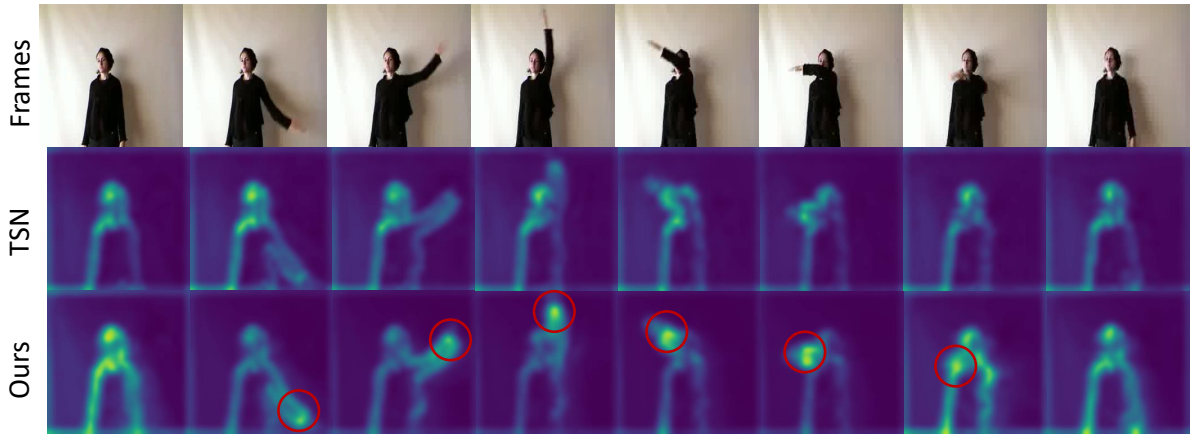

Figure 4: The visualization of input frames, attention maps from TSN and the proposed model. The attention maps are computed as square of mean values along channels. We mark the attention region with red circle for more intuitive comparison.

Mothed	Kernel(k)	Dilation(d)	Acc. (%)	Speed
TSN	-	-	36.15	54.34
TSN+MKTB	3	1,2,3	50.40	44.14
TSN+MKTB	3	1,2,3,5	53.10	43.49
TSN+MKTB	3	1	47.18	45.76
TSN+MKTB	1,3	1	47.42	45.05
TSN+MKTB	1,3,5	1	49.67	44.19
TSN+MKTB	1,3,5,7	1	53.10	43.47
TSN+1D-Conv	3	1	38.36	53.56
TSN+GRB	1	1	41.58	51.62

Table 7: Experiments of different instantiations of MKTB and GRB. A single kernel size (k) or dilation rate (d) indicates that only one branch exists. Evaluations are performed on validation set of IsoGD.

achieve the multi-scale modeling. The second section in Table 7 shows the results of temporal modeling with different kernel sizes and dilation rates. It is shown that both MKTB module with four branches and four dilation rates achieve the best performance. Essentially, both methods have similar motivation, *i.e.*, making use of convolutions with different receptive field . Compared with the single scale temporal features (*i.e.* kernel size = 3), multi-scale temporal features achieve much better performance, demonstrating that the MKTB modules capture short-term and long-term temporal dependencies effectively, thus improving the results. To verify the effectiveness of global information, we conduct experiment that replaces GRB module with 1D temporal convolution. The third section in Table 7 shows that the proposed GRB module performs better than ordinary 1D convolutions, demonstrating that GRB provides effective global information based on cross-channel similarity.

Comparison with deeper 3DCNNs. Although shallow 3DCNNs such as Res3D[Miao *et al.*, 2017] have been studied, here we further explore deeper 3DCNN on gesture recognition task. Table 3 shows that although action recognition methods P3D and I3D capture spatiotemporal information by 3D convolutions, they still perform worse than on gesture recognition task. We hypothesize that they fail to model the long-term temporal dependencies which are important for gesture recognition. In contrast, we model both the short-term and long-term temporal information via multi-kernel temporal convolutions and global refinement.

4.6 Visualization of feature maps

We qualitatively verify the ability of temporal modeling by visualizing the feature maps from intermediate layer ($res3$ in this experiment). As shown in Figure. 4, the most significant difference between the proposed model and TSN is that our model is able to attend to regions with hand in a temporally consistent way. We highlight the region by red circles in the visualization. Note that these regions are always around the hand. In contrast, TSN model overlooks the attention information in parts of the frames. The qualitative comparison further proves the effectiveness of the proposed method.

5 Conclusion

In this paper, we propose an effective and efficient learning paradigm for gesture recognition based on 2DCNNs. Two simple and plug-and-play modules are introduced, including MKTB and GRB. MKTB models the pyramidal temporal features at local-level with 1D depthwise convolutions and GRB captures the global temporal features based on the cross-channel similarity. Quantitative and qualitative results on two large scale benchmark datasets show that our model is superior to existing methods. Extensive experiments on video understanding task and video-based person re-identification task further demonstrate the effectiveness of the proposed modules. For the future work, since MKTB and GRB can be plugged into existing architectures without introducing much overhead, we will further investigate the effectiveness of these two modules on more video understanding tasks .

References

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Goyal *et al.*, 2017] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, volume 2, page 8, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Kopuklu *et al.*, 2018] Okan Kopuklu, Neslihan Kose, and Gerhard Rigoll. Motion fused frames: Data level fusion strategy for hand gesture recognition. In *CVPR Workshops*, pages 2103–2111, 2018.

[Li *et al.*, 2016] Yunan Li, Qiguang Miao, Kuan Tian, Yingying Fan, Xin Xu, Rui Li, and Jianfeng Song. Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In *ICPR*, pages 25–30. IEEE, 2016.

[Lin *et al.*, 2018] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. *arXiv preprint arXiv:1811.08383*, 2018.

[Ma *et al.*, 2018] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. *arXiv preprint arXiv:1811.02146*, 2018.

[Miao *et al.*, 2017] Qiguang Miao, Yunan Li, Wanli Ouyang, Zhenxin Ma, Xin Xu, Weikang Shi, and Xiaochun Cao. Multimodal gesture recognition based on the resc3d network. In *ICCV*, pages 3047–3055, 2017.

[Narayana *et al.*, 2018] Pradyumna Narayana, Ross Beveridge, and Bruce A Draper. Gesture recognition: Focus on the hands. In *CVPR*, pages 5235–5244, 2018.

[Qiu *et al.*, 2017] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, pages 5533–5541, 2017.

[Tran *et al.*, 2018] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018.

[TwentyBn, 2017] TwentyBn. The 20bn-jester dataset v1. https://20bn.com/datasets/jester, 2017.

[Wan *et al.*, 2016] Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera, and Stan Z Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *CVPR Workshops*, pages 56–64, 2016.

[Wang *et al.*, 2016] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016.

[Wang *et al.*, 2018] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[Zhang *et al.*, 2017] Liang Zhang, Guangming Zhu, Peiyi Shen, Juan Song, Syed Afaq Shah, and Mohammed Bennamoun. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *ICCV*, pages 3120–3128, 2017.

[Zhang *et al.*, 2018] Liang Zhang, Guangming Zhu, Lin Mei, Peiyi Shen, Syed Afaq Ali Shah, and Mohammed Bennamoun. Attention in convolutional lstm for gesture recognition. In *NeurIPS*, pages 1957–1966, 2018.

[Zheng *et al.*, 2016] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884. Springer, 2016.

[Zhou *et al.*, 2018] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018.

[Zhu *et al.*, 2016] Guangming Zhu, Liang Zhang, Lin Mei, Jie Shao, Juan Song, and Peiyi Shen. Large-scale isolated gesture recognition using pyramidal 3d convolutional networks. In *ICPR*, pages 19–24. IEEE, 2016.

[Zhu *et al.*, 2017] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition. In *IJCAI*, pages 3595–3601, 2017.

[Zolfaghari *et al.*, 2018] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *ECCV*, pages 695–712, 2018.