Convolutional Non-local Spatial-Temporal Learning for Multi-Modality Action Recognition

Ziliang Ren[1], Huaqiang Yuan[1], Wenhong Wei[1], Tiezhu Zhao[1], Qieshi Zhang [2,3*],

[1] *School of Science and Technology, Dongguan University of Technology, Dongguan, China*
[2] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*
[3] *The Chinese University of Hong Kong, Hong Kong SAR, China*

Traditional deep convolutional networks (ConvNets) have shown that both RGB and depth are complementary for video action recognition. However, it is difficult to enhance the action recognition accuracy because of the limitation of the single ConvNets to extract the underlying relationship and complementary features between these two kinds of modalities. In this paper, we proposed a novel two stream ConvNet for multi-modality action recognition by joint optimization learning to extract global features from RGB and depth sequences. Specifically, a non-local multi-modality compensation block (NL-MMCB) is introduced to learn the semantic fusion features for the recognition performance. Experimental results on two multi-modality human action datasets, including NTU RGB+D 120 and PKU-MMD dataset, verify the effectiveness of our proposed recognition framework and demonstrate that the proposed NL-MMCB can learn complementary features and enhance the recognition accuracy.

Introduction: Video action is longer-term activities that span several seconds, which requires the integration of the spatial and temporal information to achieve better accuracy. In recent years, Convolutional Networks (ConvNets) have been demonstrated profound features learning ability for images classification tasks [1]. To date, many ConvNets-based approaches have been proposed to learn discriminative features for classification tasks, which have obtained competitive results in video action recognition.

Most existing ConvNets-based methods [2] learn the spatial-temporal features from raw data directly according to the sampling strategy and spatial-temporal information representation, e.g. Temporal Shift Module (TSM) [3], motion representation [4] and dynamic images [5], which significantly enhanced the recognition performance by using the entire RGB-D sequences. However, these methods are not enough to capture the contextual relationships between different frames. While 3D and recurrent ConvNets can capture temporal information and improve the accuracy of action recognition, which are computationally expensive and cover only partial frames of the entire video. Besides, some approaches have been proposed to combine RGB, optical flow, and depth stream in two-stream frameworks for multimodal-based action recognition [6].

In this paper, inspired by the development performance of the temporal shift [3] and non-local operations [7] for capturing long-range dependencies and multi-modality learning strategy, we propose a video action recognition framework that exploits non-local module and multi-modality learning to enhance the recognition ability of model. The proposed recognition framework is composed of a two streams and non-local multi-modality compensation block (NL-MMCB), which are used to extract complementary features from different streams.

Our main contributions are as follows:

- An effective two-stream ConvNet recognition framework with a non-local multi-modality compensation block to learn complementary features from different modalities.
- Extensive experimental results demonstrate that by integrating two-stream and NL-MMCB together, the proposed recognition recognition framework outperforms the existing methods on two benchmark multi-modality datasets.

Related Work: In the past two decades, the emergence of large-scale video datasets has made the ConvNets-based approaches a great success on many image and video classification tasks.

Up to now, some powerful frameworks, such as VGG, Resnet and BNInception, are proposed successively to improve the potential of ConvNets framework in image or video classification. However, these methods can not effectively obtain temporal information of actions. Therefore, some sampling and representation mothods of raw videos are proposed to obtain richer spatial-temporal information for features learning of ConvNets models, e.g. sparse sampling, motion representation and dynamic images representation, etc. Wang *et al.* [1, 10] proposed a sparse sampling strategy to model long-range temporal information over the whole video. Yang *et al.* [11] extended the 2D kernels to 3D ConvNets to learn spatial-temporal features. Furthermore, Wang *et al.* [12] employed a rank pooling function to encode the RGB video into one dynamic image (pseudo RGB image) for capturing the features from the whole action sequecne. Besides, Lin *et al.* [3] proposed an effective Temporal Shift Module (TSM) to capture temporal relationships, which have high frame rate and recongnition performance for action classification.

Compare with RGB videos, the depth sequences can provide richer geometric information of the object. Therefore, depth-based action recognition has received widespread attention. Wang *et al.* [13] constructed different dynamic images based on depth sequences as the inputs of ConvNets, and achieved the state-of-art results on three large datasets by fine-tuning the pre-trained VGG models. Moreover, Ke çeli *et al.* [14] present a method to extract deep features from 2D and 3D representations, which combined 2D and 3D ConvNets models for human action recognition.

Simonyan and Zisserman [15] proposed a two-stream ConvNets framework to learn the spatial and temporal information from still frames and multi-frame optical flow respectively, and then it obtained the-state-of-the-art results on video actions benchmark datasets. Based on the two-stream framework, some sampling strategies and compression representation methods are designed to obtain more spatiaal-temporal information for feature learning, such as sparse temporal sampling [1], motion and dynamic image representation [4, 5], etc. Furthermore, multi-modality joint learning methods are proposed to extract single- and cross-modal features for action recognition, such as c-ConvNet [12], SC-ConvNets [16], MDNN [17] and J-ResNet-CMCB [18] etc.

Proposed Framework: The proposed recognition framework first samples a pair of RGB and depth frome an RGB-D sequence, and then a novel deep ConvNet is designed to extract the features for RGB and depth modalitie, as shown in Fig. 1. This deep ConvNet includes a two-stream separate basic ConvNets to learn the RGB and depth sequences features, and the NL-MMCB is used to learn the compensation information from different modalities.

Since the remarkable performance of the non-local learning mechanism and residual network on image classification tasks, we design the proposed recognition framework based on the no-local block [7] and ResNet model [19] to extract spatial-temporal features. The proposed recognition framework includes two parallel paths, each of which contains 7×7 convolution layers and 4 bottleneck blocks. The proposed NL-MMCB is attached at the ending of residual block for fusing compensation features.

Non-Local Multi-modality Compensation Block: To reduce the number of parameters and enhance the efficiency of the recognition framework, we design non-local multi-modality compensation block, named NL-MMCB, to learn the compensation features from different modalities, as shown in Fig. 1.

Fig. 2 shows the detail of compensation learning, the NL-MMCB contains RGB and depth information streams. After feeding to the NL-MMCB, the feature maps X_{rgb} and X_d obtained from the RGB and depth streams, and the information stream can be represented as

$$F_{rgb} = F_{non-local}(X_{rgb}), \tag{1}$$

$$F_d = F_{non-local}(X_d), \tag{2}$$

where $F_{non-local}$ denotes the non-local learning operator. Based on NLB learning strategy, the two information flows can be represented as

$$\theta_r = \phi_r = g_r = F_{rgb}, \tag{3}$$

$$\theta_d = \phi_d = g_d = F_d, \tag{4}$$

where θ_r, ϕ_r and g_r are inputed to the following layers, and the compensation feature are represented as

$$Z_{rgb} = F_{non-local}(\theta_r, \phi_r, g_d) + F_{rgb}, \tag{5}$$

$$Z_d = F_{non-local}(\theta_d, \phi_d, g_r) + F_d, \tag{6}$$

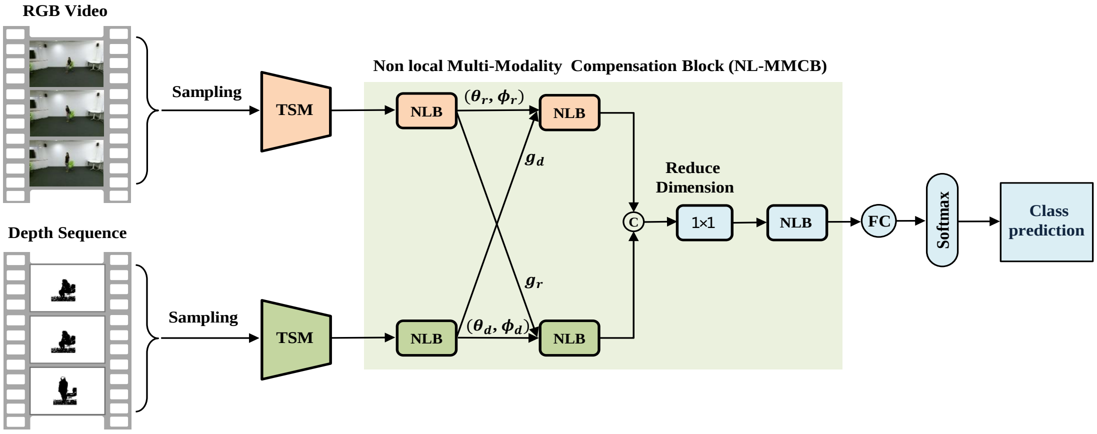
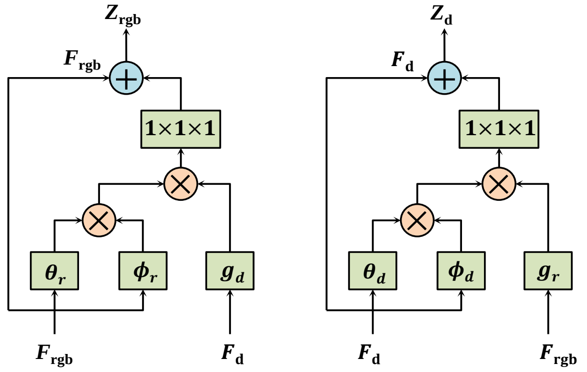


Fig. 1 *Overview of our recognition framework. Our recognition framework based on TSM and two-stream ConvNets consists of the two separate paths and multi-modality features learning module (NL-MMCB). "©" denotes the features concatenation operator and* 1×1 *is used to implement dimensionality reduction operation.*



(a) RGB information flow. (b) Depth information flow.

Fig. 2 *Non-local block with compensation learning. "*F_{rgb}*" and "*F_d*" denote the inputs of RGB and depth information flow, respectively.*

where "+F_{rgb}" and "+F_d" denote residual connection [19]. The feature maps of two streams are integrated together through concatenation operations. Then the compensation features can be obtained

$$Z_{rgb-d} = F_{non-local}(ReLU(BN(f(W_{rgb-d} * [Z_{rgb}, Z_d])))), \quad (7)$$

where W_{rgb-d} is a learnable 1×1 convolution kernel, and "$[\cdot, \cdot]$" denotes concatenation. We insert batch normalization and ReLU layers into NL-MMCB. In this way, the number of channels in Z_{rgb-d} is same to Z_{rgb} and Z_d, e.g., $7 \times 7 \times 2048$ for 2048 channels after conv5x, and we can obtain the number of channels is 2048 for Z_{rgb}, Z_d and Z_{rgb-d}.

Joint Optimization: A softmax layers is placed on top of the fully connected NL-MMCB, and estimate the class probability score by

$$prob_c = \frac{exp(W_c X_c + b_c)}{\sum_{c_i} exp(w_{c_i} X_c + b_{c_i})}, \quad (8)$$

where W_c and b_c are the weight and bias, respectively. The loss function is expressed as

$$L(y, C) = -\sum_{c=1}^{C} y_c(log(prob_c)), \quad (9)$$

where C is the number of action classes, and $prob_c$ is the class probability score.

Implementation Details: In the experiments, we first use a sampling strategy to obtain the T pairs (RGB, depth). According to the need of recognition framework, we set the spatial size is 224×224 pixels, and feed obtained T pairs (RGB, depth) to the proposed framework. Our recognition framework is initialized by using the pretrained weights of ResNet-50 in [19]. In addition, as suggested by [10, 1], the dropout technique and minibatch stochastic gradient descent with momentum are employed to enhance training efficiency, and the batch size, momentum and initial learning rate are set to 64 , 0.8 and 0.001, respectively. The same sampling strategy and discretization method are used to botain T pairs of images, and the softmax scores are used to predict the classification of the action.

Datasets: The NTU RGB+D 120 [8] is the largest benchmark action recognition dataset, which is an extension of NTU RGB+D 60. In this dataset, it includs more variation of environmental conditions, more subjects and camera views, etc. It consists of 114,880 multi-modality samples collected from 106 different subjects, and there are 120 action types in the dataset. Same as 60 dataset, it provides two kindes of protocols for the proposed recognition framework evaluation, including C-Sub and C-Set. There are 63,360 training and 51,120 testing samples in C-Sub protocol, while 54,720 training and 59,760 testing samples in C-View protocol, respectively.

PKU-MMD [9] is the largest continuous multi-modality 3D human action dataset. It consists of 1076 long video sequences. The dataset contains 51 action categories and provides four kinds of modalities for action detection and recognition, including four different modalities, RGB, depth, infrared radiation and skeleton. The 1,076 long sequences can be divided into 20,734 human action with well annotation information, and the dataset provides two kinds of evaluation protocols, cross-subject (C-Sub) and cross-view (C-View). There are 18,134 training samples and 2,600 testing samples in C-Sub protocol, while 13,847 training and 6,887 testing samples in C-View protocol, respectively.

Experiments: : The effect of the proposed action framework (based model: ResNet-50) is first verified on NTU RGB+D 120 dataset, and Table 1 shows the results. According to the experimental results in references [1, 10], the segment T is set from 1 to 8 to evaluate the recognition performance of the proposed framework, and the proposed method with NL-MMCB based on (RGB, depth) achieved an accuracy of 87.8% and 89.2% for C-Sub and C-Set protocols, respectively. The experimental results illustrate the effectiveness of the proposed recognition framework. Experimental results also show that increasing the value of segment numbers T will improve the recognition performance of our model. However, the recognition performance cannot incease when the segment number T increases from 6 to 8. Therefore, to strike a balance between recognition performance and computational burden, the segment number T is set to 6 in following experiment.

Table 1: Comparative accuracy of different T on the NTU RGB+D 120 dataset. Based model: ResNet-50. Notation: D: Depth.

Segment Numbers	Modality	C-Sub	C-Set
$T=1$	(RGB,D)	56.5%	56.9%
$T=3$	(RGB,D)	84.12%	83.8%
$T=6$	(RGB,D)	87.8%	89.2%
$T=8$	(RGB,D)	87.4%	89.3%

Further experiments are conducted to test ResNet-50, ResNet-50 with temporal shift module [3] and non local module [7], and our proposed action recognition framework. We recognize human action by fusing RGB and depth modalities for base model ResNet-50, and joint training RGB and depth modalities for our proposed framework, and Table 2 shows the results. The results of SC-ConvNets are 86.8% and 88.1% for C-Sub and C-Set protocols with dynamic images when segment is set to 3, respectively, and the accuracy is boosted to 87.8% and 89.2% with (VDIs, DDIs) for our method with the proposed NL-MMCB. In specially, the proposed recognition ConvNets without NL-MMCB will degenerate to the plain SC-ConvNets as in [6]. It indicate that the proposed NL-MMCB can learn complementary features from RGB and depth modalities, and the complementary features will improve the recognition accuracy.

Table 2: Comparative accuracy of our method and previous approaches on the NTU RGB+D 120 dataset. Segment T is set to 6 in TSM and our models, and based model: ResNet-50. Notation: D: Depth, OF: Optical flow, S: Skeleton.

Method	Modality	C-Sub	C-Set
ResNet-50 [19]	RGB	40.6%	39.5%
ResNet-50 [19]	D	43.6%	42.2%
ResNet-50 [19]	RGB + D	51.4%	51.3%
TSM + non-local [3]	RGB	86.8%	88.1%
TSM + non-local [3]	D	85.8%	86.3%
TSM + non-local [3]	RGB + D	87.0%	88.6%
two-stream network [8]	RGB + D + OF + S	64.0%	66.1%
J-ResNet-CMCB [18]	(RGB, D)	82.8%	83.6%
SC-Convnets [6]	(RGB, D)	86.8%	88.1%
Our method	(RGB, D)	87.8%	89.2%

Further experiments are conducted on multi-modality human action dataset PKU-MMD, and the comparison results are shown in Table 2. The proposed recognition framework with NL-MMCB achieves the 92.2% and 93.1% on C-Sub and C-View protocols. The recognition accuracy of our method outperform the results [21] by margin 5.3% and 0.5%, respectively, which are comparable to the state-of-the-art [6].

Table 3: Comparison of our recognition framework to the State-of-the-Art on the PKU-MMD dataset [20].

Method	Modality	C-Sub	C-View
LSTM [21]	S	83.7%	91.0%
SA-LSTM [21]	S	86.3%	91.4%
TA-LSTM [21]	S	86.6%	92.3%
STA-LSTM [21]	S	86.9%	92.6%
J-ResNet-CMCB [18]	(RGB, D)	90.4%	91.4%
SC-ConvNets (ResNet-101) [6]	(RGB, D)	92.1%	93.2%
Our method (ResNet-50)	(RGB, D)	92.2%	93.1%

Conclusion: In this paper, a novel two-stream ConvNets with NL-MMCB is presented for multi-modality action recognition. The proposed recognition framework employed a non local compensation block and joint optimization strategy to learn complementary features from differnet modalities. The NL-MMCB is designed to extract the complementary features from different modalities. Experimental results demonstrated the effectiveness of the proposed recognition framework. Compared with the existing approaches, the proposed method obtained comparable results for two benchmark multi-modality datasets.

Acknowledgment: This work was supported by the Dongguan Science and Technology of Social Development Program (No.20211800904722), Special Project in Key Fields of Guangdong Colleges and Universities (No.2021ZDZX3007) and National Natural Science Foundation of China (No. 61872083). *Corresponding author: Qieshi Zhang (qs.zhang@siat.ac.cn).

References

1 L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 11, pp. 2740–2755, 2019.

2 B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with deeply transferred motion vector CNNs," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 5, pp. 2326–2339, 2018.

3 J. Lin, C. Gan, K. Wang, and S. Han, "TSM: Temporal shift module for efficient and scalable video understanding on edge devices," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–15, 2020.

4 E. Ijjina and K. Chalavadi, "Human action recognition in RGB-D videos using motion sequence information and deep learning," *Pattern Recognition*, vol. 72, pp. 504–516, 2017.

5 B. Fernando, E. Gavves, J. Oramas M., A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 4, pp. 773–787, 2017.

6 Z. Ren, Q. Zhang, J. Cheng, F. Hao, and X. Gao, "Segment spatial-temporal representation and cooperative learning of convolution neural networks for multimodal-based action recognition," *Neurocomputing*, vol. 433, pp. 142–153, 2021.

7 X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

8 J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L. Duan, and A. Kot Chichung, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 42, no. 10, pp. 2684–2701, 2020.

9 C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," *CoRR*, vol. abs/1703.07475, 2017.

10 L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Gool, "Temporal segment networks: Towards good practices for deep action recognition," *European Conference on Computer Vision (ECCV)*, vol. 9912, pp. 20–36, 2016.

11 C. Li, Y. Hou, P. Wang, and W. Li, "Multiview-based 3D action recognition using deep networks," *IEEE Transactions on Human-Machine Systems (THMS)*, vol. 49, no. 1, pp. 95–104, 2019.

12 P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, "Cooperative training of deep aggregation networks for RGB-D action recognition," *32nd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 7404–7411, 2018.

13 P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Ogunbona, "Depth pooling based large-scale 3D action recognition with convolutional neural networks," *IEEE Transactions on Multimedia (TMM)*, vol. 20, no. 5, pp. 1051–1061, 2018.

14 A. Keçeli, A. Kaya, and A. Can, "Combining 2D and 3D deep models for action recognition with depth information," *Signal, Image and Video Processing*, vol. 12, pp. 1197–1205, 2018.

15 K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems (NIPS)*, vol. 1, no. 1, pp. 568–576, 2014.

16 Z. Ren, Q. Zhang, J. Cheng, F. Hao, and X. Gao, "Segment spatial-temporal representation and cooperative learning of convolution neural networks for multimodal-based action recognition," *Neurocomputing*, vol. 433, pp. 142–153, 2021.

17 Y. Tang, Z. Wang, J. Lu, J. Feng, and J. Zhou, "Multi-stream deep neural networks for RGB-D egocentric action recognition," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 29, no. 10, pp. 3001–3015, 2019.

18 J. Cheng, Z. Ren, Q. Zhang, X. Gao, and F. Hao, "Cross-modality compensation convolutional neural networks for RGB-D action recognition," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, pp. 1–12, 2021.

19 K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

20 J. Hu, W. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 11, pp. 2186–2200, 2017.

21 S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based LSTM networks for 3D action recognition and detection," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 7, pp. 3459–3471, 2018.