Rethinking Fusion Baselines for Multi-modal Human Action Recognition

Hongda Jiang, Yanghao Li, Sijie Song, and Jiaying Liu(✉)

Institute of Computer Science and Technology, Peking University, Beijing, China
{jianghd,lyttonhao,ssj940920,liujiaying}@pku.edu.cn

Abstract. In this paper we study fusion baselines for multi-modal action recognition. Our work explores different strategies for multiple stream fusion. First, we consider the early fusion which fuses the different modal inputs by directly stacking them along the channel dimension. Second, we analyze the late fusion scheme of fusing the scores from different modal streams. Then, the middle fusion scheme in different aggregation stages is explored. Besides, a modal transformation module is developed to adaptively exploit the complementary information from various modal data. We give comprehensive analysis of fusion schemes described above through experimental results and hope our work could benefit the community in multi-modal action recognition.

Keywords: Fusion · Multi-modality · Action recognition

1 Introduction

With the rapid development of deep learning, there has been tremendous progress in computer vision [9,11,13]. The two-stream network [17] makes remarkable contribution for action recognition, by fusing the results of spatial and temporal streams, and achieves good performance on popular action recognition benchmarks. However, it still remains confusing whether combining different modalities on the final results as two-stream is the best choice. Recently, Spatiotemporal Multiplier Networks [5] investigate the middle connections in the two-stream architecture. The connections in their work are straightforward and their method only considers RGB and optical flow data. With the development of depth cameras, depth and other modal data become more available, which are able to provide clues for action recognition from other perspectives [14]. However, there is a lack of exploration of generic multi-modal fusion schemes. Thus, in our paper, we rethink various fusion schemes and design sufficient experiments to give some insights in the multi-modal fusion.

We conduct a general research on the fusion baselines for multi-modal human action recognition. We consider aggregating modalities in different levels, *i.e.*,

This work was supported by National Natural Science Foundation of China under contract No. 61772043 and CCF-Tencent Open Research Fund.

R. Hong et al. (Eds.): PCM 2018, LNCS 11166, pp. 178–187, 2018.
https://doi.org/10.1007/978-3-030-00764-5_17

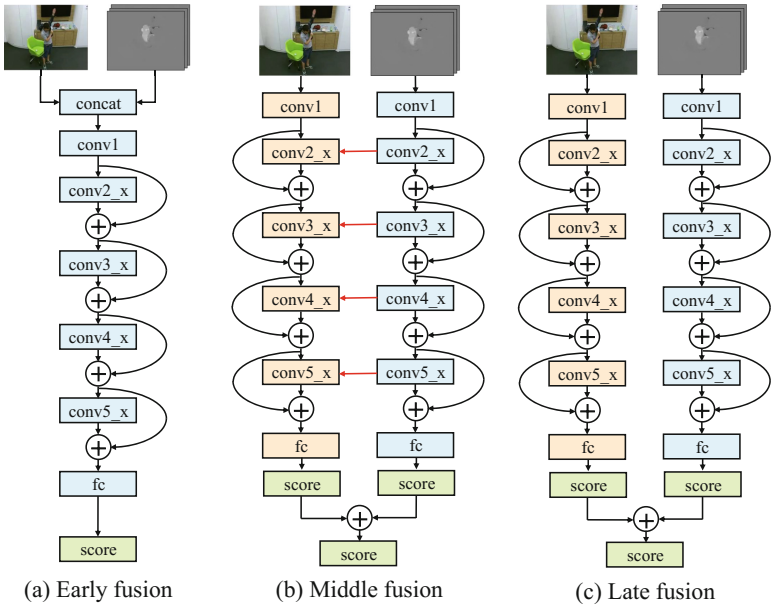


Fig. 1. The architectures of different multi-modal fusion schemes with ResNet101 as backbone (From left to right, early fusion, middle fusion and late fusion). Note that the input modalities are not limited in the above two modalities. We will explore more different modalities in the following.

early fusion, middle fusion and late fusion. In the early fusion, we stack different modalies together as a single input which is the most straightforward and easiest way to fuse different modalities. Next, we explore the late fusion scheme which directly fuses the softmax scores of different modal streams. Finally, the middle fusion is presented which combines multi-modal data in the feature level. We systematically explore various stages of middle fusion and propose a modal transformation module in adaptive middle fusion. Our networks for different fusion methods are based on ResNet101 [7] and the architectures of our different fusion schemes are illustrated in Fig. 1.

Our contributions are listed as follows:

- We conduct a thorough investigation on various fusion baselines which aggregate multiple modalities in different levels.
- We adopt deep ablation analysis of different fusion stages for middle fusion method. Sufficient experiments are conducted and discussed to compare different fusion methods with multiple modalities.
- We further propose a novel modal transformation module for middle fusion method, leading to a more efficient model combination over existing simple middle fusion methods.

2 Related Work

Pioneer video-based action recognition research mainly focuses on crafting features from videos, such as Motion Boundary Histograms [2], Histograms Of Flow [12], subsequent Dense Trajectories [22], and Improved Dense Trajectories [23].

Since videos are sequential data which contain plenty of temporal information [10], the key point of video action recognition is how to model temporal dynamics. Derived from Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) network has the ability to capture long-term and short-term information [6,20]. It is natural to employ LSTM to model these sequential data [3,15,19]. Meanwhile, CNN architectures are also proved useful in the video-related task. The C3D network [8,21] is widely applied because of its 3D convolution, which can simultaneously catch spatial and temporal information.

The most relative work to ours is two-stream [17] structure, which parallelly processes the spatial and temporal streams and fuses their prediction scores. Lately, Spatiotemporal Residual Networks [4] extend the original two-stream approach by building middle connections. To further understand how the interaction works, Spatiotemporal Multiplier Networks [5] provide a systematic investigation on the middle fusion in residual connections based on ResNet50 and ResNet152 [7]. In contrast to previous efforts which only consider RGB and optical flow, we take a step further and conduct a comprehensive exploration on various multi-modal fusion schemes, taking RGB, optical flow and depth information into account. Besides, a modal transformation module is proposed to achieve a more efficient modal combination.

3 Exploring Different Multi-modal Fusion Schemes

In this section, we investigate different fusion schemes for action recognition. We fuse the multi-modal data from input level, score level and feature level, respectively, which corresponds to early fusion, late fusion and mid fusion.

We use ResNet101 [7] as our backbone. ResNet101 is a fully convolutional architecture, with a chain of residual units. Each residual unit consists of closely linked 1×1 and 3×3 convolutions and is equipped with additive skip connections. In the end of the ResNet101 there is an average pooling and a fully connected layer. Fusions take place at different parts of the networks.

3.1 Early Fusion

Early fusion suggests that we fuse the multiple modalities on the input by stacking them along the channel dimension. Assume there are M kinds of modalities for action recognition and the input for the i-th modality is $\mathbf{I}_i (i = 0, ..., M-1)$. Note that $\mathbf{I}_i$ for different modalities should have the same spatial resolution but may differ in the number of channels. We concatenate the different modal data as:

$$\mathbf{I} = \mathbf{I}_0 \oplus \mathbf{I}_1 \oplus \mathbf{I}_2 \oplus ... \oplus \mathbf{I}_i \oplus ... \oplus \mathbf{I}_{M-2} \oplus \mathbf{I}_{M-1}, \tag{1}$$

where $\mathbf{I}$ is the final input for early fusion, $\oplus$ indicates concatenation along channels. Early fusion is the most straightforward and comprehensive method to combine multiple modalities. However, the multi-modal data get mixed in a low-level manner. It might be hard for the network to extract discriminative features for action recognition.

3.2 Late Fusion

Late fusion fuses the scores of different modal streams. Similar to two-stream, the softmax scores from multiple streams are combined together by average fusion. Supposing M modalities are adopted to classify N actions and the score of the i-th modality is $\mathbf{p}_i = (p_{i,0}, p_{i,1}, ..., p_{i,N-1})$. The final score $\mathbf{y} = (y_0, y_1, ..., y_{N-1})$ and the prediction label z can be given as below:

$$y_j = \frac{1}{M} \sum_{k=0}^{M-1} p_{k,j}, \tag{2}$$

$$z = \underset{j}{\operatorname{argmax}}(y_j), \tag{3}$$

where y_j indicates the score of the j-th action class ($j = 0, 1, ..., N-1$).

3.3 Multiple Middle Fusion

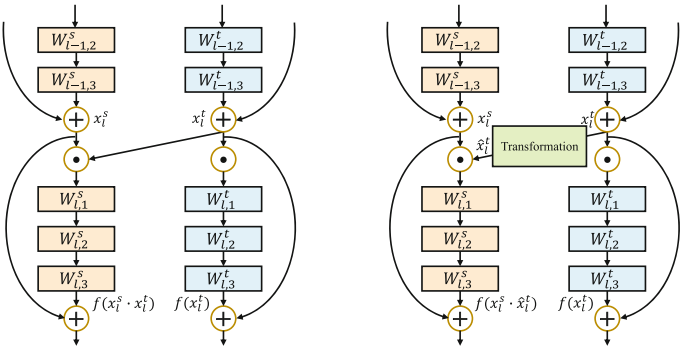
Direct and Adaptive Connections. For simplicity, we explain the middle fusion with two modalities, while more modalities could be introduced for middle fusion. As illustrated in Fig. 1, we design our middle fusion networks by building adaptive connections between different streams. A detailed schematic is proposed as (b) in Fig. 2. $W^s_{l,c}$ indicates the weights of the c-th convolution layer in the l-th residual unit and s, t represent different streams. We formulize the proposed method as:

$$\widehat{x}^t_l = g(x^t_l), \tag{4}$$

$$\widehat{x}^s_{l+1} = x^s_l + f(x^s_l \cdot \widehat{x}^t_l), \tag{5}$$

where x^s_l, x^t_l are features from the l-th layer of the two streams, respectively, and $f(\cdot)$ is the residual function. In addition, we insert a general transformation, $g(\cdot)$, into each connection which transforms the features adaptively for a more sufficient fusion. As a special case, the Spatiotemporal Multiplier Networks [5] directly fuse two modalities in which the $g(\cdot)$ is an identity transformation and we call it direct connection in Fig. 2. The final results are obtained by averaging the scores of two streams.

The middle fusion is more complicated when extending to multiple modalities since the middle connections is directional. The number of schemes increases when including new modalities and we design our connections according to the involved modalities. An example will be given in Sect. 4.



(a) Direct connection (b) Adaptive connection

Fig. 2. An illustration of middle fusion. (a) Direct connection represents identity transformation connection and (b) adaptive connection corresponds to variable transformation connection.

4 Experiments

4.1 Dataset and Settings

In this section, we evaluate the performance of the above methods with NTU RGB+D Dataset [16] (NTU). NTU is a multi-modal dataset which is captured indoor by Microsoft Kinect v2 cameras concurrently. It consists of 56,880 action samples containing aligned RGB videos, depth map sequences, 3D skeleton data and infrared videos. We adopt RGB, optical flow and depth videos for our multi-modal action recognition. We split the videos into 40,400 training samples and 16,480 testing samples following the cross-subject rule in [16].

Implementation Details. We use ResNet101 [7] as our backbone network structure and follow the same training schemes in Temporal Segment Networks (TSN) [24] architecture. During training, the input images are first resized to 256×340 and then cropped with the specific width and height which are randomly chosen from $\{256, 224, 192, 168\}$, followed by resizing to 224×224. The cropping is performed on the four corners or the center of images. Since the input channels of different modalities may differ, we follow the initialization method in [24] to use models pretrained on ImageNet for all modalities and then modify the weights in the first convolution layer. In the training process, we randomly select 3 segments for each video where each segment contains one RGB image, one depth image or stacked optical flow frames. To speed up the testing process, we average the results of 3 segments for each video to obtain the final results. Note that when comparing to the state-of-the-art methods, we average the results of 25 segments for a fair comparison.

For middle fusion, we proceed in two steps during training. We first independently train each modality and then insert our adaptive connections among

different streams. Finally, the connected networks with different modalities are optimized jointly with cross entropy losses. We propose the adaptive connection in Sect. 3. In the experiments, we apply a 1×1 convolution layer as the adaptive transformation which is expected to lead to a more efficient fusion.

Table 1 shows the classification results with a single modality. This demonstrates the importance of motion information in action recognition. The performance from different modalities also reveals that different modalities may have complementary information and the fusion could achieve better results.

Table 1. Results with a single modality on the NTU dataset in accuracy (%).

	RGB	Flow	Depth
Acc. (%)	83.87	**92.02**	85.23

4.2 Middle Fusion Stage Exploration

Since ResNet101 has four stages of convolution blocks, we conduct an exploration experiment to investigate where to append the middle connections. For each stage, ResNet101 has multiple residual units and we link the second residual unit for middle connection. We first compare the results of appending one middle connection in each stage, respectively, and then apply four connections for all stages. Here we apply the experiment based on the direct connections.

The results are in Table 2. With the fusion stage changing from 1 to 4, the result is continuously increasing and we achieve the best performance when appending all connections. This indicates that our middle fusion method could benefit from different levels of feature fusions. Therefore, in the following experiments, we apply connections in all four stages.

Table 2. Results for different stages of connections in middle fusion with direct connections on the NTU dataset.

Stage1	Stage2	Stage3	Stage4	Acc. (%)
√	-	-	-	93.83
-	√	-	-	93.93
-	-	√	-	94.07
-	-	-	√	94.29
√	√	√	√	**94.37**

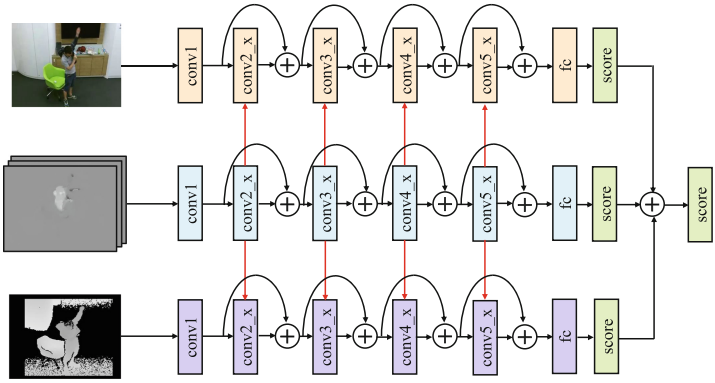


Fig. 3. Connections for middle fusion with RGB, optical flow and depth data.

4.3 Comparisons for Different Fusion Methods

Fusion of Multiple Modalities. It is easy to extend early fusion and late fusion to more modalities. However, the connections are directional for middle fusion. For two modalities, we follow the connection direction in [5] to construct connections from optical flow to RGB. For three modalities, we append adaptive connections from optical flow to RGB and depth at the same time, as illustrated in Fig. 3.

We conduct fusion experiments among all the combinations and fusion methods above. The results are summarized in Table 3. From the results, we can find that:

1. The results of early fusion are even worse for those with single modal data. For example, the combination of RGB and optical flow by early fusion is inferior to optical flow. We attribute this to the high complexity of the early fusion input. It is hard for the network to extract discriminative features for the task of action recognition.
2. Comparing the direct connection and the adaptive connection for middle fusion, the latter has better performance among all the combinations. This demonstrates that our proposed model is able to achieve a more sufficient fusion by transforming the features adaptively.
3. Experimental results show that late fusion performs better than direct connections. It is mainly because the middle fusion network with direct connections suffers from overfitting. Nevertheless, our middle fusion by adaptive connections achieve better performance than late fusion, illustrating the effectiveness of our modal transformation module.
4. The combination of depth and optical flow is even superior to the three modalities. We ascribe the confusing result to the considerable noise as we only select 3 segments for the test. Once we average the results of 25 segments, our model achieves the best performance with the three modalities in Table 4.

Table 3. Results of different fusion methods with multi-modal data on the NTU dataset in accuracy (%).

RGB	Flow	Depth	Early	Late	Direct connections	Adaptive connections
√	√	-	89.77	94.45	94.37	**94.62**
√	-	√	80.63	87.99	87.16	**88.14**
-	√	√	89.29	94.98	94.47	**95.03**
√	√	√	82.49	94.98	94.53	**94.98**

4.4 Comparisons with the State-of-the-Art Methods

We compare our method with current state-of-the-art models. For a fair comparison, we mark the modalities employed in each method. Table 4 shows that our method outperforms other methods using the same or fewer modalities. With the three modalities, our model achieves the best result.

Table 4. Comparisons with state-of-the-art methods on the NTU dataset in accuracy (%).

Methods	Skeleton	RGB	Flow	Depth	Acc. (%)
STA-LSTM [18]	√	-	-	-	73.4
VA-LSTM [25]	√	-	-	-	79.4
P-CNN [1]	-	√	√	-	53.8
TSN (BN-Inception) [24]	-	√	√	-	88.5
Chained MT [26]	√	√	√	-	80.8
Late fusion	-	√	√	-	94.8
Late fusion	-	√	√	√	95.2
Adaptive connections	-	√	√	-	**95.2**
Adaptive connections	-	√	√	√	**95.5**

5 Conclusion

In this paper, we investigate different fusion baselines for multi-modal action recognition. We explore early fusion, middle fusion and late fusion, respectively, which aggregate the multi-modal data from different levels. A modal transformation module is proposed to help effectively utilize the complementary information and improve the action recognition results. Analysis shows that our modal transformation module with adaptive connections has the best performance among all the fusion methods. We hope the insights from this work could encourage further research in multi-modal action recognition.

References

1. Chéron, G., Laptev, I., Schmid, C.: P-CNN: pose-based CNN features for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3218–3226 (2015)
2. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Proceedings of European Conference on Computer Vision, pp. 428–441 (2006)
3. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
4. Feichtenhofer, C., Pinz, A., Wildes, R.: Spatiotemporal residual networks for video action recognition. In: Proceedings of Advances in Neural Information Processing Systems, pp. 3468–3476 (2016)
5. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 7445–7454 (2017)
6. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: continual prediction with LSTM. Neural Comput. **12**(10), 2451–2471 (2000)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
9. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of ACM International Conference on Multimedia, pp. 675–678 (2014)
10. Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High dynamic range video. ACM Trans. Graph. **22**, 319–325 (2003)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
12. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
13. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
14. Liu, J., Li, Y., Song, S., Xing, J., Lan, C., Zeng, W.: Multi-modality multi-task recurrent neural network for online action detection. IEEE Trans. Circ. Syst. Video Technol. (2018)
15. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3D human action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 816–833. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_50
16. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: a large scale dataset for 3D human activity analysis. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
17. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Proceedings of Advances in Neural Information Processing Systems, pp. 568–576 (2014)
18. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: AAAI, vol. 1, p. 7 (2017)

19. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. IEEE Trans. Image Process. **27**(7), 3459–3471 (2018)
20. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using LSTMs. In: Proceedings of International Conference on Machine Learning, pp. 843–852 (2015)
21. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
22. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3169–3176 (2011)
23. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of IEEE International Conference on Computer Vision, pp. 3551–3558 (2013)
24. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Proceedings of European Conference on Computer Vision, pp. 20–36 (2016)
25. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2117–2126 (2017)
26. Zolfaghari, M., Oliveira, G.L., Sedaghat, N., Brox, T.: Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In: Proceedings of IEEE International Conference on Computer Vision, pp. 2923–2932 (2017)