SKELETON-INDEXED DEEP MULTI-MODAL FEATURE LEARNING FOR HIGH PERFORMANCE HUMAN ACTION RECOGNITION

Sijie Song [1], *Cuiling Lan* [2*], *Junliang Xing* [3], *Wenjun Zeng* [2], *Jiaying Liu* [1*]

[1] Institute of Computer Science and Technology, Peking University, Beijing, China
[2] Microsoft Research Asia, Beijing, China
[3] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China

ABSTRACT

This paper presents a new framework for action recognition with multi-modal data. A skeleton-indexed feature learning procedure is developed to further exploit the detailed local features from RGB and optical flow videos. In particular, the proposed framework is built based on a deep Convolutional Network (ConvNet) and a Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM). A skeleton-indexed transform layer is designed to automatically extract visual features around key joints, and a part-aggregated pooling is developed to uniformly regulate the visual features from different body parts and actors. Besides, several fusion schemes are explored to take advantage of multi-modal data. The proposed deep architecture is end-to-end trainable and can better incorporate different modalities to learn effective feature representations. Quantitative experiment results on two datasets, the NTU RGB+D dataset and the MSR dataset, demonstrate the excellent performance of our scheme over other state-of-the-arts. To our knowledge, the performance obtained by the proposed framework is currently the best on the challenging NTU RGB+D dataset.

Index Terms— Multi-modal, skeleton-indexed, high performance, action recognition

1. INTRODUCTION

Human action recognition is an important task in video analytics. It benefits a wide range of applications such as intelligent video surveillance, human-computer interaction. However, it is challenging due to cluttered backgrounds, varied illumination, camera motion, and subtle differences of some action categories, especially when interacting with objects. With the prevalence of the RGB-D cameras and the advance of pose estimation techniques, multi-modal data (*e.g.*, RGB, optical flow, skeleton) can provide complementary cues and be utilized to enhance the recognition performance.

In the past decades, many pioneering works have extensively studied human action recognition from RGB videos. Compared with the hand crafting of feature representations [1], ConvNet-based methods [2, 3] are able to learn features automatically from raw data and mainly focus on the combination of RGB images and optical flow to utilize both spatial and temporal contexts. Some works [4, 5] attempt to model the long-term temporal dynamics using RNN. Most of those works generally extract features in a global manner, and may lose local detailed information of interest. However, when perceiving the real-word, a human usually has a global observation and next focuses on local parts to get more detailed information. In addition, for distinguishing fine-grained actions (*e.g.*, *brush teeth* and *apply a lipstick*), the interacted objects are helpful for differentiating some action types. Thus, local features in specific regions provide compensation for missing details [6, 7]. Peng *et al.* [6] integrate a region of interest (ROI) pooling layer to improve action representations. However, the output of ROI pooling corresponds to a wide receptive field and fails to describe local details. Cheron *et al.* [7] extract part patches according to the postures obtained from pose estimation. They take an SVM as the classifier and the network is not end-to-end trainable, limiting the recognition capability. Besides, the temporal dynamics are not effectively explored. In this work, we study how to explore both the local and global information for efficient human action recognition in an end-to-end network with temporal dynamics explored, based on skeleton data and RGB videos.

Skeleton, as a high-level human representation, is robust to the variation of illumination, backgrounds and viewpoints. Skeleton data has been an attractive option for action recognition [8, 9]. Recent works leverage recurrent neural network architectures [8, 9] to automatically learn the effective feature representations. The lack of appearance information results in the ambiguity of some actions whenever only using skeleton data, such as *play with phone* and *type on a keyboard*. In contrast, the RGB appearance information can eliminate such

[*] Corresponding author. The majority of the work was done at Microsoft Research Asia. This work was supported by National Natural Science Foundation of China under contract No. 61772043 and No. 61672519, Microsoft Research Asia fund under project ID FY17-RES-THEME-013, and CCF-Tencent Open Research Fund. We also acknowledge the support of NVIDIA Corporation with the GPU for this research.

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

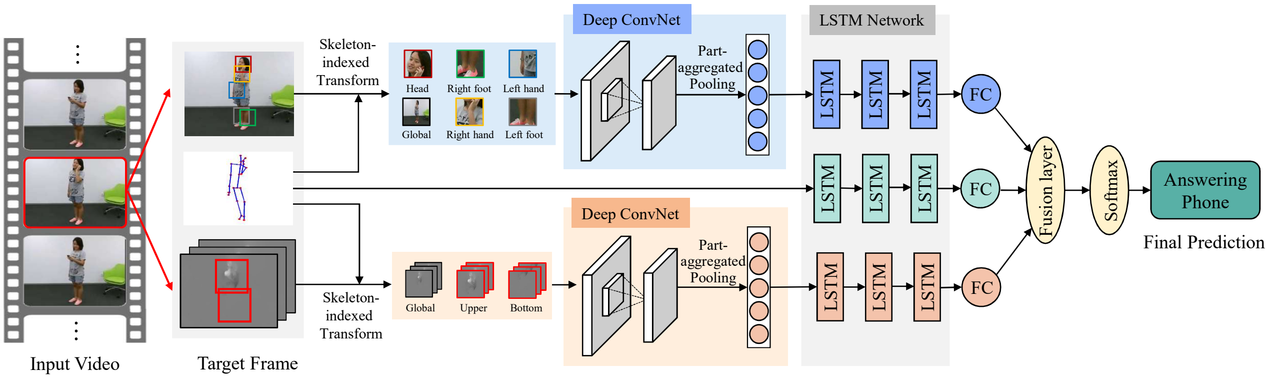


Fig. 1. The framework of the proposed skeleton-based multi-modal deep feature learning method which is integrated with a skeleton-indexed transform layer and a part-aggregated pooling layer. Here we show the late fusion scheme when combining different modalities.

uncertainty. Therefore, it is worth combing multi-modal data to make use of the complementary information from them.

In this paper, we propose a unified network with skeleton-indexed feature learning to integrate multi-modal data for action recognition, as shown in Fig. 1. With skeleton joints aligned to RGB frames, local image patches around key points are extracted with our skeleton-indexed transform layer, in which local details with high resolution are explicitly captured. Since the dimension of skeleton-indexed features depends on the number of actors within one frame, we employ a part-aggregated pooling layer to adapt to different scenarios. Deep convolutional networks are utilized to extract spatial features for RGB and optical flow data. Stacked LSTM layers are constructed for each modal data to model long-term temporal dynamics more efficiently. Finally, we investigate several fusion schemes to take advantage of different modalities.

The rest of the paper is organized as follows: Sec. 2 describes our skeleton-based multi-modal feature learning method. Sec. 3 shows the effectiveness of our proposed method through experiment results. Finally, concluding remarks are given in Sec. 4.

2. PROPOSED METHOD

In this section, we describe our proposed framework with skeleton-indexed multi-modal feature learning in details. The overall structure is shown in Fig. 1. It consists of three streams. Guided by the skeleton sequence, the RGB (or optical flow) stream goes through skeleton-indexed deep ConvNet for frame-wise feature learning, and the LSTM network for temporal dynamic exploring of the features. Skeleton sequences go through an LSTM network for action recognition. The features from the three streams are fused for final action classification. Given a video sequence $\mathbf{V} = \{V_t : t = 1, ..., T\}$, we extract skeleton-indexed spatial features from each RGB frame or from each optical flow stack by a deep ConvNet with a skeleton-indexed transform layer. Then, a part-aggregated pooling layer regulates the features to a fixed length. LSTM is followed to further explore the temporal information and give the final prediction.

2.1. Skeleton-Indexed Transform Layer

Features extracted from an image is successful in representing the overall structure. But it is challenging for them to describe subtle variations, such as the differences of objects the actor interacts with, or the local small movement . To capture such local details, we design a spatial transform layer to crop and resize the local region based on the key joints. In this work, we leverage the key joints of skeleton data (which can be easily obtained by a Kinect-like camera or pose estimation technologies) to select the local regions, both in RGB and optical flow.

For RGB frames, given $(x_{t,k,p}, y_{t,k,p})$, which is the coordinate of the p-th joint on the k-th actor in frame V_t, by the transform layer, the patch $V_{t,k,p}$ of part p is cropped with a ratio r:

$$V_{t,k,p} = \mathcal{T}(V_t, x_{t,k,p}, y_{t,k,p}, r), \tag{1}$$

where $\mathcal{T}(\cdot)$ denotes the projecting, cropping and resizing operation. Note the images are cropped as squares, with the side length of each patch being $2r$. A resizing of the cropped square to a fixed size (*e.g.*, 224 × 224) is performed to enlarge the local details. Then we get an image set $\{V_{t,1,1}, V_{t,1,2}, ...V_{t,k,p}, ...V_{t,K,P}\} \cup V_t$, where K denotes the number of actors and P denotes the number of parts of interest within one frame. In our case, we use the five parts (head, left hand, right hand, left foot and right foot) of each actor in the RGB frames, *i.e.*, $P = 5$. Note, in order to preserve the global spatial structure of each part, we also include the original image V_t into the image set.

For optical flow, we crop the upper and bottom body parts, *i.e.*, $P = 2$, as shown in Fig. 1. Given all the joint coordinates $\{(x_{t,k,p}, y_{t,k,p})\}_{p=1}^{J}$ of actor k, we can locate the boundary

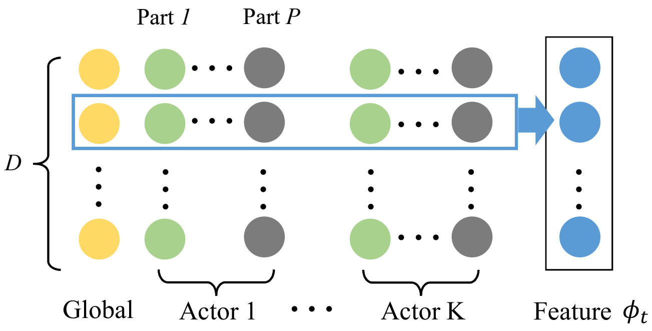


Fig. 2. Part-aggregated pooling layer, which generates a fixed length output regardless of the actor numbers.

of the actor by $(x_{t,k}^{min}, x_{t,k}^{max}, y_{t,k}^{min}, y_{t,k}^{max})$. The center of the upper body part is $((x_{t,k}^{min}+x_{t,k}^{max})/2, (3y_{t,k}^{min}+y_{t,k}^{max})/4)$ and that of the lower body part is $((x_{t,k}^{min}+x_{t,k}^{max})/2, (y_{t,k}^{min}+3y_{t,k}^{max})/4)$ if the upper-left is set as the coordinate origin. Given the part centers, the patch of each body part can be obtained per Eq. (1). Correspondingly, we obtain the image set for the optical flow stream.

Each image patch in the corresponding image set is fed into the following convolutional layers to encode spatial features. In practice, we can take the popular ConvNet structure, *e.g.*, VGG [10], BN-Inception [3], as our backbone.

2.2. Part-Aggregated Pooling Layer

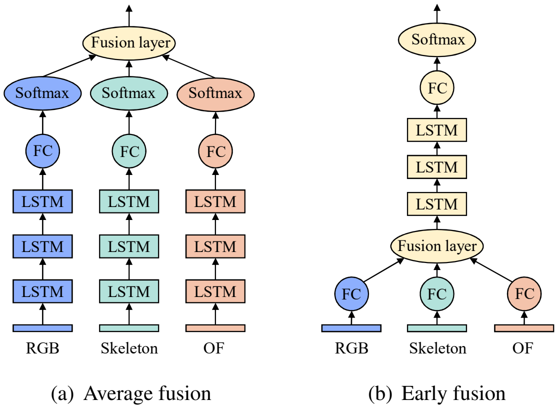
Each patch can be encoded to a feature vector of D dimension through the ConvNet as described above. Then the dimension of the features we get is $(K \times P + 1) \times D$. The number of actors K is unfixed which depends on the action categories (*e.g.*, *shake hands* for pairs and *drink* for single). For the purpose of further classification, it is necessary to design a method to generate a fixed length output regardless of the value of K. Here we employ a part-aggregated pooling layer to deal with this issue. As shown in Fig. 2, the features from different elements are arranged to a map. Then our part-aggregated pooling is performed as:

$$\phi_t = (\phi_{t,g} + \sum_{k=1}^{K}\sum_{p=1}^{P}\phi_{t,k,p})/(KP+1), \tag{2}$$

where $\phi_{t,k,p}$ is the convolutional feature for the part p of actor k in frame V_t, and $\phi_{t,g}$ is the global representation for frame V_t. The fixed-length outputs $\{\phi_t : t = 1, ..., T\}$ are then passed into a recurrent neural network for temporal dynamic modelling.

2.3. Temporal Dynamic Modelling

To exploit the temporal dependencies of the features in the video sequence, LSTM [11] is employed to build our network. As shown in Fig. 1, we build the network by staking several LSTM layers followed by a fully connected layer for classification.



(a) Average fusion (b) Early fusion

Fig. 3. Average and early fusion schemes.

For a sequence, the scores for C classes are given as:

$$\mathbf{z} = \sum_{t=1}^{T}(\mathbf{W}\mathbf{h}_t + \mathbf{b}), \tag{3}$$

where $\mathbf{z} = [z_1, z_2, ..., z_C]^{\mathrm{T}}$, T denotes the length of the video sequence, $\mathbf{h}_t$ is the hidden state of the top LSTM layer which is then fed into a fully connected layer with learned weights $\mathbf{W}$ and $\mathbf{b}$. Note, $\mathbf{W}$ and $\mathbf{b}$ are shared parameters over all time steps. Thus, the predicted probability of the i-th class can be computed as:

$$p\left(c_i|X\right) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, \quad i = 1, ..., C. \tag{4}$$

2.4. Multi-Modal Fusion

With skeleton, RGB and optical flow data, we have three streams in our network. Different modality has different characteristics and they are complementary. To utilize complementary information from these modalities, we design three schemes to fuse the multiple streams at different levels with corresponding loss functions, including average fusion, early fusion, and late fusion.

Average Fusion: Average fusion is a simple yet effective way to fuse different streams and leverage the complementary strengths. In average fusion as shown in Fig. 3(a), each stream is treated as a separate task with the same classification label. During training, the class scores of each task are used to compute a loss value. Then the final loss function is defined as the sum of each individual loss, which is given as:

$$\mathcal{L} = \sum_{m\in\{s,r,o\}} \mathcal{L}^m = -\sum_{m\in\{s,r,o\}}\sum_{i=1}^{C} y_i \log \hat{y}_i^m, \tag{5}$$

where $\mathbf{y} = (y_1, \cdots, y_C)^{\mathrm{T}}$ denotes the groundtruth and $y_i \in \{0, 1\}$, $\hat{y}_i^m$ indicates the probability that the sequence is predicted as the i-th class from modal m. During testing, the class scores of all tasks are averaged to form the final prediction of the action class. We also try to learn the weighted combination in an end-to-end manner. But we do not observe much significant improvement.

Early Fusion: This scheme fuses the three streams at an early stage, as shown in Fig. 3(b), after the feature extraction module of RGB and optical flow images. Given the feature vectors from RGB, optical flow and skeleton streams of the t-th frame, ϕ_t^r, ϕ_t^o and ϕ_t^s, we first map them with fully connected layers to the vectors $f(\phi_t^r)$, $f(\phi_t^o)$, $f(\phi_t^s) \in \mathbb{R}^{N\times 1}$ and then concatenate them to $[f(\phi_t^r)^{\mathrm{T}}; f(\phi_t^o)^{\mathrm{T}}; f(\phi_t^s)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{N\times 3}$. The fusion layer fuses the features as:

$$\phi_t' = [f(\phi_t^r)^{\mathrm{T}}; f(\phi_t^o)^{\mathrm{T}}; f(\phi_t^s)^{\mathrm{T}}]^{\mathrm{T}} * \mathbf{w}_e, \qquad (6)$$

where $\mathbf{w}_e \in \mathbb{R}^{3\times 1}$ is the fusion matrix and can be learned automatically. Then the following recurrent module takes the fused features $\{{\phi'}_t : t = 1, ..., T\}$ as input. Cross-entropy loss is used as the objective function.

Late Fusion: The features from the top layer of LSTM modules are high-level representations. We adopt a late fusion scheme to make use of the information from each modality. As shown in Fig. 1, we add a fusion layer before *softmax*. For each stream, the features from the top LSTM layer are summarized as Eq. (3) and we get $\mathbf{z}^r$, $\mathbf{z}^o$, $\mathbf{z}^s \in \mathbb{R}^{C\times 1}$. In the fusion layer, we combine different streams as $[{\mathbf{z}^r}^{\mathrm{T}}; {\mathbf{z}^o}^{\mathrm{T}}; {\mathbf{z}^s}^{\mathrm{T}}]^{\mathrm{T}} * \mathbf{w}_l$. The network automatically learns the weight $\mathbf{w}_l \in \mathbb{R}^{3\times 1}$ with the goal of minimizing cross entropy loss.

3. EXPERIMENT RESULTS

3.1. Datasets and Settings

To evaluate the effectiveness of our multi-modal framework, we conduct experiments on the following datasets: the NTU RGB+D Dataset [12], which is so far the largest multi-modal dataset for action recognition, and the MSR 3D Daily Activity Dataset [13], designed for human-object interactions. We also show that with the help of pose estimation technology, our method can be applied when 3D skeleton is not provided, which can be found in the supplemental material.

NTU RGB+D Dataset (NTU) [12] is the largest action recognition dataset with multi-modal data (skeleton, RGB, depth and IR). This dataset consists of 56880 video samples with more than 4 million frames. There are 60 action types performed by 40 subjects, including interactions with pairs and individual activities. The cross subject (CS) and cross view (CV) settings are two protocols to evaluate the performance. Because the resolution of RGB frames is 1920×1080 and the background is empty, after resized to 224×224, the actors only occupy a small area in the scene. To avoid unnecessary degradation in performance, the original RGB images are cropped to increase the resolution of subjects. In addition, we randomly select videos to form a NTU subset under cross-view protocol with 6000 videos for training and 3065 for testing. Note that to accelerate the extraction of optical flow [3], the RGB videos and skeleton data are downsampled with a stride of 5.

MSR 3D Daily Activity Dataset (MSR) [13] is a daily activity dataset with 16 action categories, including skeleton (20 joints for each skeleton), RGB and depth. The total number of video sequences is 320 with image resolution of 640×480. The first five actors are used for training and others for testing, which is a cross-subject split. The high intra-class variation and the small scale of this dataset make the recognition on the dataset challenging.

Implementation Details: For RGB and optical flow, we take the BN Inception network as utilized in [3] as our backbone ConvNet for feature extraction. Each frame is finally represented by a vector of dimension 1024 (*i.e.*, $D = 1024$) from the *global pool* layer of BN Inception. For the NTU dataset, we employ a 3-layer LSTM network, using 1024 neurons in the LSTM for RGB/optical flow streams and 150 neurons for skeleton. For the MSR dataset, to alleviate overfitting on such a small dataset, we employ a 1-layer LSTM network, using 100 units in the LSTM for RGB/optical flow. The number of neurons in the skeleton stream is set to be 50. The batch sizes for the NTU and MSR datasets are 256 and 16, respectively. Adam [14] is adopted to automatically adjust the learning rate with an initial value of 0.001. Dropout with a probability of 0.5 is used to mitigate overfitting.

3.2. Effectiveness of Skeleton-Indexed Features

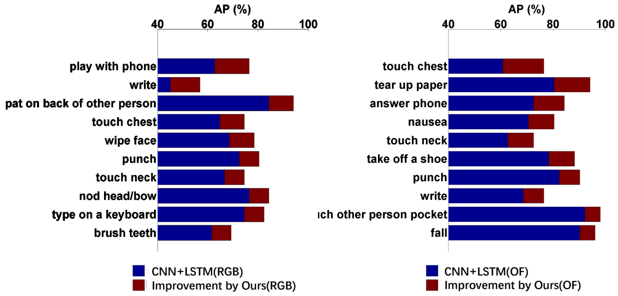
We first evaluate the proposed method with single modal data in Table 1. CNN+LSTM(RGB) and CNN+LSTM(OF) denote the baselines of RGB and optical flow, respectively, which directly feed the global convolutional features into the LSTM network. For the NTU dataset, compared with the results with single modal data in TSN [3], which are denoted as TSN(RGB) and TSN(OF), the application of LSTM brings about 7%-12% gain for RGB and 2%-6% for optical flow, benefiting from the merits of LSTM which is capable of exploring temporal long range dynamics. For the MSR dataset, the result of RGB stream is improved by about 4% owing to the temporal modelling, though the performance of optical flow degrades due to overfitting. Note that TSN [3] is currently one of the best networks for RGB-based action recognition. We train that network using their source code to get the baseline performance.

In our method, the proposed skeleton-indexed local convolutional features are employed, followed by LSTM layers. We find that for optical flow, the performance is improved by about 1%-2% over CNN+LSTM on the NTU dataset, but 5% on the MSR dataset. This is mainly because many actions are motionless in the NTU dataset. For RGB data, there is about 1% gain over CNN+LSTM in all the dataset settings, thanks to the capability of capturing of action details.

Besides, we list the top 10 activity categories in the NTU subset in Fig. 4, for which the recognition performance is improved the most by using the proposed skeleton-indexed features in our framework. We see that for RGB, the most improved action types are usually interactions with objects, such

Table 1. Evaluation of the effectiveness of our skeleton-indexed feature learning scheme in accuracy (%).

Methods	NTU subset	NTU-CS	NTU-CV	MSR
TSN(RGB) [3]	69.95	74.34	76.35	64.37
CNN+LSTM(RGB)	78.99	81.83	88.53	68.13
Ours(RGB)	**80.75**	**82.05**	**89.96**	**69.37**
TSN(OF) [3]	80.55	85.15	87.17	77.50
CNN+LSTM(OF)	82.77	87.60	93.33	72.50
Ours(OF)	**84.11**	**88.83**	**94.15**	**78.13**



(a) RGB

(b) Optical Flow

Fig. 4. Top 10 action categories in the NTU subset for which the recognition performance is improved the most by using the proposed skeleton-indexed features.

as *play with phone* and *write*. Meanwhile, motion-related actions are most improved in the stream of optical flow, such as *touch chest*, *tear up paper*. This confirms that our skeleton-indexed features are better in describing the detailed information and the temporal dynamic exploration is helpful.

3.3. Combination of Different Modalities

The performance of different kinds of modalities and their combination are investigated with results shown in Table 2. In this subsection, we adopt the average fusion scheme on different modal branches when testing. Note, the results of RGB and optical flow streams are derived from our proposed skeleton-indexed features with LSTM.

From Table 2, it is observed that optical flow usually dominates the final performance, from which we conclude that motion information is one of the key factors for action recognition. In our experiments, the combination of different modal data significantly improves the results over single modality, due to the complementary information provided by each stream. Though the performance from skeleton is inferior to RGB and optical flow, it still contributes to the fusion results, since 3D skeleton data can provide additional depth information and high level information with background excluded. The highest performance is achieved by fusing the three streams. The gain of recognition with multi-modal data over single modality is about 3% - 10%, and even reaches 12% for the MSR dataset, owing to the complementary information of different modalities.

Table 2. Recognition with multi-modal data in accuracy(%).

Methods	NTU subset	NTU-CS	NTU-CV	MSR
Ske.	68.22	72.54	83.53	70.00
RGB	80.75	81.83	89.96	69.37
OF	84.11	88.83	94.15	78.13
Ske. & RGB	83.30	85.12	92.82	85.63
Ske. & OF	85.97	90.19	95.77	80.00
RGB & OF	88.78	90.65	96.33	84.83
Ske. & RGB & OF	**90.15**	**91.40**	**97.15**	**90.63**

3.4. Effects of Fusion Schemes

In this subsection, we evaluate the performance of different fusion schemes in Table 3, including average fusion, early fusion and late fusion. In the late fusion, we initialize the network with the parameters trained from average fusion, then freeze all the layers except the fusion layer. We see that the late fusion brings about 1% improvement compared with average fusion and achieves the best performance, since the network learns the combination of high-level representations automatically. We also observe that the learned weight of optical flow is the highest, indicating that the network is most confident with the scores from optical flow. However, for the early fusion, the results are not satisfactory. It is probably because the features from different modalities mix in an early stage, and the training procedure of the network can be easily dominated by the components from one modality and get trapped to sub-optimal results.

Table 3. Results of different fusion schemes in accuracy(%).

Methods	NTU subset	NTU-CS	NTU-CV	MSR
Average Fusion	90.15	91.40	97.15	90.63
Early Fusion	85.84	90.42	94.87	84.38
Late Fusion	**91.62**	**92.55**	**97.90**	**91.88**

3.5. Comparison with Other State-of-the-Arts

Table 4. Performance comparisons on the NTU in accuracy (%).

Methods	Ske.	RGB	OF	CS	CV
Trust Gate [15]	✓			69.2	77.7
STA-LSTM [8]	✓			73.4	81.2
P-CNN [7]		✓	✓	53.80	61.68
TSN [3]		✓	✓	88.48	90.41
Chained-MS [16]	✓	✓	✓	80.8	–
SI-MM (Ours)		✓	✓	90.65	96.33
SI-MM (Ours)	✓	✓	✓	**92.55**	**97.90**

Table 4 and Table 5 show the performance comparisons of our skeleton-indexed multi-modal framework (SI-MM) with other state-of-the-art approaches for the NTU and MSR datasets, respectively. Note that we implement the P-CNN [7] with the pretrained CNN model provided by the authors, and the SVM with linear kernel is adopted. For a fair comparison, we list the involved data modalities of each method. Thanks to the skeleton-indexed local features, the semantic contexts provide effective cues for recognition. Besides, the application of LSTM makes it possible to leverage long-term

Table 5. Performance comparisons on the MSR in accuracy (%).

Methods	Ske.	RGB	OF	Acc. (%)
Action Ensemble [13]	✓			68.0
Moving Poselets [17]	✓			74.5
P-CNN [7]		✓	✓	61.88
TSN [3]		✓	✓	88.12
SI-MM (Ours)		✓	✓	84.38
SI-MM (Ours)	✓	✓	✓	**91.88**

temporal dynamics. We can see that our SI-MM outperforms other state-of-the-art methods, especially on the NTU dataset, which outperforms TSN [3] by 6% and Chained-MS [16] by 12% with the same modal inputs. Note, only the cross subject result is provided in [16]. Limited by the tiny scale of the MSR dataset, it is hard to train the network due to overfitting, but we still get comparable performance.

4. CONCLUSIONS

In this paper, we propose a skeleton-indexed multi-modal feature learning framework for high performance action recognition. Since semantic cues, local information, and temporal dynamics are all important for action classification, we design our network to explore both global and local features guided by the key joints of skeletons, with temporal dynamics explored by the LSTM network. We employ a part-aggregated layer to solve the unfixed length of skeleton-indexed features. To better incorporate different modalities and take advantage of their complementary information, we explore different fusion schemes of the RGB, optical flow and skeleton streams. Experiments on two benchmarks demonstrate the effectiveness of our framework.

5. REFERENCES

[1] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013, pp. 3551–3558.

[2] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014, pp. 568–576.

[3] 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*, 2016, pp. 20–36.

[4] Zuxuan Wu, Yu-Gang Jiang, Xi Wang, Hao Ye, and Xiangyang Xue, "Multi-stream multi-class fusion of deep networks for video classification," in *ACM MM*, 2016, pp. 791–800.

[5] Jiaying Liu, Yanghao Li, Sijie Song, Junliang Xing, Cuiling Lan, and Wenjun Zeng, "Multi-modality multi-task recurrent neural network for online action detection," *Trans. on Circuit System for Video Technology*, vol. PP, no. 99, pp. 1–1, 2018.

[6] Xiaojiang Peng and Cordelia Schmid, "Multi-region two-stream r-cnn for action detection," in *ECCV*, 2016, pp. 744–759.

[7] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid, "P-CNN: Pose-based cnn features for action recognition," in *ICCV*, 2015, pp. 3218–3226.

[8] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *AAAI*, 2017, pp. 4263–4270.

[9] Yueyu Hu, Chunhui Liu, Yanghao Li, Sijie Song, and Jiaying Liu, "Temporal perceptive network for skeleton-based action recognition," in *BMVC*, 2017.

[10] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016, pp. 1010–1019.

[13] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012, pp. 1290–1297.

[14] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[15] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *ECCV*, 2016, pp. 816–833.

[16] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *ICCV*, 2017, pp. 2923–2932.

[17] Lingling Tao and René Vidal, "Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition," in *ICCVW*, 2015, pp. 61–69.