IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 July 2022, accepted 19 August 2022, date of publication 24 August 2022, date of current version 6 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3201227

RESEARCH ARTICLE

Spatial-Temporal Information Aggregation and Cross-Modality Interactive Learning for RGB-D-Based Human Action Recognition

QIN CHENG[1,2], **ZHEN LIU**[2], **ZILIANG REN**[3], **JUN CHENG**[2], **(Member, IEEE), AND JIANMING LIU**[4]

[1]School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China
[2]CAS Key Laboratory of Human–Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[3]School of Science and Technology, Dongguan University of Technology, Dongguan 523808, China
[4]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China

Corresponding author: Jianming Liu (liujm_guet@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant U1913202, Grant U1813205, Grant 61872083, and Grant 61262074; in part by the Chinese Academy of Sciences (CAS) Key Technology Talent Program, Shenzhen Technology Project under Grant JCYJ20180507182610734; and in part by the Dongguan Science and Technology of Social Development Program under Grant 20211800904722.

ABSTRACT The RGB-D-based human action recognition is gaining increasing attention because the different modalities can provide complementary information. However, the recognition performance is still not satisfactory due to the limited ability to learn spatial-temporal feature and insufficient inter-model interaction. In this paper, we propose a novel approach for RGB-D human action recognition by aggregating spatial-temporal information and implementing cross-modality interactive learning. Firstly, a spatial-temporal information aggregation module (STIAM) is proposed to utilizes sample convolutional neural networks (CNNs) to aggregate the spatial-temporal information in entire RGB-D sequence into lightweight representations efficiently. This allows the model to extract richer spatial-temporal features with limited extra memory and computational cost. Secondly, a cross-modality interactive module (CMIM) is proposed to fully fuse the multi-modal complementary information. Moreover, a multi-modal interactive network (MMINet) is constructed for RGB-D-based action recognition by embeding the above two modules into the two-stream CNNs. In order to verify the universality of our approach, two backbones are deployed in the two-stream architecture, successively. Ablation experiments demonstrate that the proposed STIAM can bring significant improvement in recognizing actions. CMIM can further play the advantages of complementary features of multiple modalities. Extensive experiments on NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD datasets proved the effectiveness of the proposed approach. The proposed approach achieves an accuracy of 94.3% and 96.5% for cross-subject and cross-view on NTU RGB+D 60, 91.7% and 92.6% for cross-subject and cross-setup on NTU RGB+D 120, 93.6% and 94.2% for cross-subject and cross-view on PKU-MMD datasets, which are the state-of-the-art performance. Further analysis denotes that our approach has advantages in recognizing subtle actions.

INDEX TERMS Deep learning, action recognition, cross-modality interactive learning, information aggregation.

I. INTRODUCTION

benefitting from the complementary relationship in multi-modal data, human action recognition based on RGB-D sequence has attracted much attention in recent years. There are two crucial ingredients for RGB-D-based action

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

recognition: extracting discriminative spatial-temporal features for each modality and fully exploring the complementary relationship between multi-modal information. For extracting spatial-temporal features, early approaches mainly focus on designing handcraft features, such as Histogram of Optical Flow (HOF) [1], Motion Boundary Histograms (MBH) [2], etc. Recent years, deep learning-based methods have shown strong advantages in accuracy and robustness, such as TSN [3], TSM [4], and TEA [5]. In order to fuse the complementary information in multiple visual modalities, several multitask networks [6], [7], [8], [9], [10], [11] for jointly training with multiple modalities have been proposed. Although deep learning-based approaches have achieved great success in action recognition, they still have shortcomings in two aspects: feature extraction of subtle motion and multi-modal fusion.

1) Early method usually training the deep network on a single frame or short clip of the entire action, which leads to insufficient modeling in time dimension. To alleviate the problem, Wang *et al.* proposed a novel Temporal Segment Networks (TSN) [3] to establish the long-term aggregation of motion information. Most of the subsequent methods [4], [5], [12] followed the segmented sampling strategy in TSN. However, due to the limitation of memory and computation cost, it is not feasible to use densely sampled frames for training, which unfortunately led to insufficient modeling of subtle actions. Alternatively, some work has designed a series of hand-made motion representations to centralize more motion information, such as Motion history image (MHI) [13] and dynamic image based on rankpooling [14], [15]. But they are not friendly to real-time scenarios because of the amount of extra computational cost.

2) The mainstream multi-modal action recognition methods can be divided into late fusion and intermediate fusion. Due to the simplicity of operation and the lack of consideration for different spatial dimensions, late fusion once dominated. Representative methods include Multimodal Correlative Representation Learning (MCRL) [6], 5-stream ConvNets [7], and Multimodal Training / Unimodal Testing(MTUT) [8]. However, these methods usually train each modality individually and fuse their recognition scores, which does not adequately learn the complementary information of different modalities. Research in neuroscience and machine learning has shown that intermediate feature fusion can benefit learning. Several works [6], [15], [16], [17] have constructed cooperative networks to implement intermediate fusion. However, the information interaction of multiple modalities in these methods is only performed at a single feature level, which is not sufficient for integrating the complementary features of multiple modalities to participate in decision-making.

To overcome the shortcomings above, we propose a novel multi-modal interaction network (MMINet), which extracts rich spatial-temporal features in subtle motion and enhances multi-modal feature interaction to improve the human action recognition performance. The main contributions of this paper are as follows:

1) A spatial-temporal information aggregation module (STIAM) is proposed to aggregate the motion information of each moment of the action into lightweight representations by utilizing sample convolutional neural networks (CNNs). This enables the model to efficiently leverage densely sampled frames for training and avoids the missing of crucial motion information.
2) A cross-modality interactive module (CMIM) is proposed to enhance information exchange between different modalities, which enables the features from different layers of multiple modalities to participate in recalibrating the channel level features at the decision level of the network. The module facilitates the model to fully learn the complementary information of different modalities and greatly improves the recognition performance.
3) We construct an end-to-end trainable multi-modal interaction network (MMINet), which can effectively utilize densely sampled RGB-D sequences for action recognition. Extensive experiments on three datasets verified the universality and effectiveness of our approach, and achieved state-of-the-art performance.

II. RELATED WORKS

In the past two decades, many researchers have devoted themselves to the study of action recognition and made great contributions. In this section, we briefly review related work from two aspects: spatio-temporal feature extraction and multi-modal fusion.

A. SPATIAL-TEMPORAL FEATURE EXTRACTION

Extracting spatial-temporal features from RGB-D sequences is an important step in action recognition. Early works before the deep learning era usually establish hand-designed features to perform action recognition, which include the histogram of optical flow (HOF) [18], histogram of oriented gradients (HOG) [19], motion boundary histograms (MBH) [20], and dense trajectories (DT/iDT) [21], etc. Nowadays, CNN-based methods have become mainstream and achieved great success, which can automatically extract features from raw data and provide high-level semantic information. Since the 2D CNN cannot explicitly learn the temporal information in the video, some works adopted sparse temporal sampling strategies [5], [22], [23] or explored temporal dependencies between video frames at multiple time scales [24], [25], [26]. Compared with 2D CNN, 3D CNN-based methods [27], [28], [29], [30] are able to directly extract spatial-temporal features from videos. Although these works have achieved good performance, the use of 3D CNN is limited by its parameters and computational overhead, prompting the emergence of the works on 3D convolutional kernel factorization [31], [32], [33] to balance the accuracy and model cost.

Another approach to extract spatial-temporal feature is to construct new representations from input data. Early works such as MEI and MHI [34] modeled the motion information of the entire video into a human action representation, where the consecutive frame differences and motion history are encoded in a single static image, respectively. Considering many actions have temporal ordering characteristics intuitively, Fernando *et al.* [14] proposed a new video representation that captures time-varying information by training a linear ranking machine in chronological order of videos. Bilen *et al.* [35] compute ranking machine directly at pixel-level of images or features extracted from CNN, which is called approximate rank pooling (ARP). The dynamic images constructed by this method encode the spatial-temporal information and enable the use of existing ConvNets models directly. Optical flow is a conventional and effective motion representation to describe movement. However, its usage is restricted due to expensive computation costs and huge storage demands [36]. Inspired by the definition of optical flow, Sun *et al.* [37] proposed a novel optical flow guided feature (OFF) representation extracted from RGB frames, which can boost the performance of action recognition with fewer computation costs than optical flow. Wang *et al.* [38] proposed to encode the RGB-D sequences based on scene flow into one motion map, called Scene Flow to Action Map(SFAM), modeling long term spatio-temporal feature for action recognition.

B. MULTI-MODAL FUSION

As a paradigm for processing multi-modal data in action recognition, the two-stream architecture was first proposed in [39] to process the spatial and temporal information in the RGB frames and optical flow respectively, and then combined the results of the two separate recognition streams by late fusion. Attribute to the great success of the two-stream architecture, multi-modal action recognition has attracted great attention in recent years. Wang *et al.* [40] treated the two-stream ConvNets as generic feature extractors and conduct trajectory-constrained pooling to aggregate convolutional features into effective descriptors. Feichtenhofer *et al.* [41] utilize the two-stream ConvNets to implement different fusion strategies for multi-modal spatial-temporal feature fusion and proposed multiplicative gating functions to build cross-stream residual connections between the two streams [42]. By inflating filters and pooling kernels of 2D ConvNets into 3D, Carreira *et al.* [28] proposed a Two-Stream Inflated 3D ConvNet (I3D) to directly extract the spatial-temporal information of the action.

Due to the availability of accurate and low-cost sensors, more and more modalities are used in action recognition [7], [43], [44], [45] How to combine these complementary modalities to generate robust and accurate recognition results has been receiving continuous attention. Zhao *et al.* [44] proposed a two-stream RNN/CNN structure to extract temporal features and spatial-temporal features from skeleton data and RGB frames respectively. Wang *et al.* [38] proposed to extract Scene Flow to Action Map (SFAM) representations jointly from RGB-D data and fed them into different ConvNets to perform the late fusion. Hu *et al.* [45] proposed to adopt bilinear pooling layers to compact RGB-D and skeleton features extracted from different networks. Wang *et al.* [46] and Ren *et al.* [16] improved the action recognition performance of RGB-D videos by training two-stream ConvNets as a single network cooperatively. Das *et al.* [47] used skeleton as an auxiliary modality to guide the RGB cues to generate spatial embeddings to better exploit both modalities for action recognition. Moreover, audio can also used to assist action recognition. Gao *et al.* [48] use audio as an efficient preview to select useful moments in untrimmed videos. Nagrani *et al.* [49] explored the correlation and obtained weak labels for action recognition between actions and the speech of characters from movie screenplays.

Recent years, cross-modality feature learning has attracted much attention due to its capability to mine relations between different modalities. Song *et al.* [50] regard skeleton as an accessorial modality to achieve cross-modality adaptive representation learning with RGB and optical flow modalities. Cheng *et al.* [15] proposed a cross-modality compensation block to learn complementary information between two modalities and compensate the unimodal features for better action recognition performance. Some works also perform cross-modality interactive feature learning at multiple levels in the network hierarchy for multi-modal fusion applications, such as Multimodal Transfer Module (MMTM) [51] and Information Aggregation-Distribution Module (IADM) [52]. Moreover, audio is usually converted to a spectral representation to assist action recognition [53], [54] and the multi-modal fusion is performed at the feature level. In addition to action recognition, sevral cross-modality interaction strategies are also proposed in some person re-identification (ReID) works. The two-stream architectures is often adopted to handle modality-aware collaborative learning through partial parameter sharing [55], [56] or eliminate the large modality gap through well designed loss functions [57].

III. THE PROPOSED APPROACH

A. THE FRAMEWORK OF OUR APPROACH

As shown in Fig. 1, the framework of our approach consists of two parts: spatial-temporal information aggregation and multi-modal interaction learning.

The video frames and depth sequence are first divided into k segments, respectively. Each segment is fed to the STIAM to encode into an aggregation data with the same size as one video frame. Then, the k pairs of RGB-D aggregated data are fed to backbone networks with interactive modules for deeper feature extraction. The CMIM connects the features at different levels of the two backbone networks, prompting the two network branches to jointly optimize and learn complementary information from each other. Unlike previous method based on handcrafted action descriptors [13], [14], we do not need to calculate and store aggregated data in

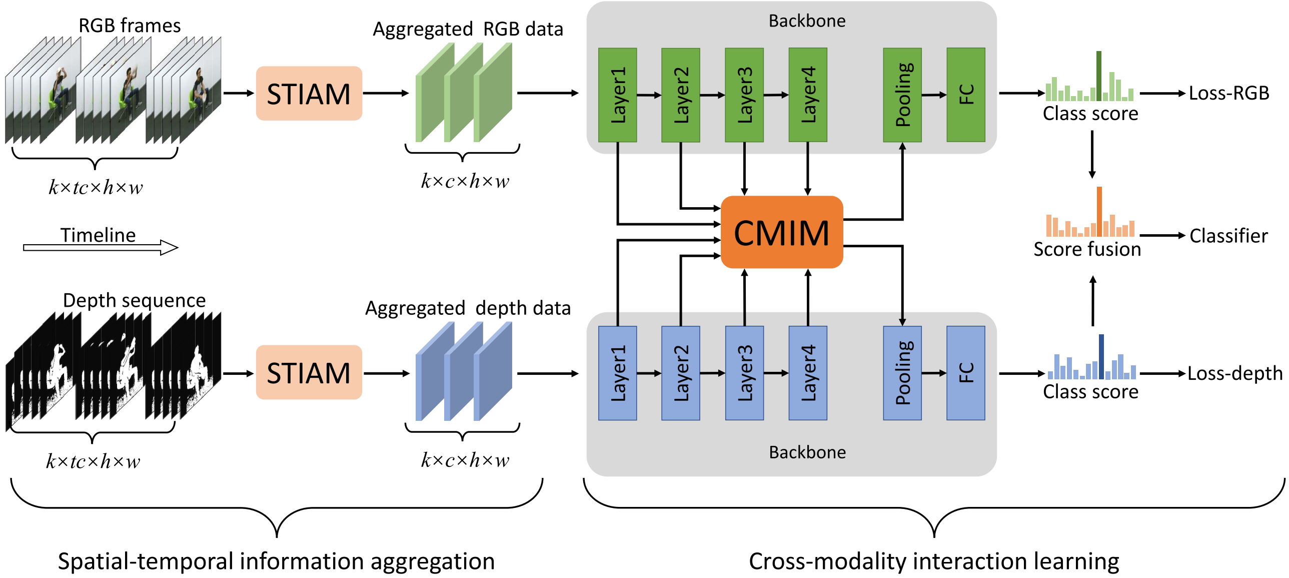

FIGURE 1. Overview of the proposed MMINet, which consists of two components: spatial-temporal information aggregation and cross-modality interaction learning. The RGB-D sequences are first encoded into pairs of aggregated data by STIAM to capture comprehensive motion information, and CMIM is used to fully exchange the complementary information of different modalities. It is worth mentioning that the entire network is end-to-end trainable.

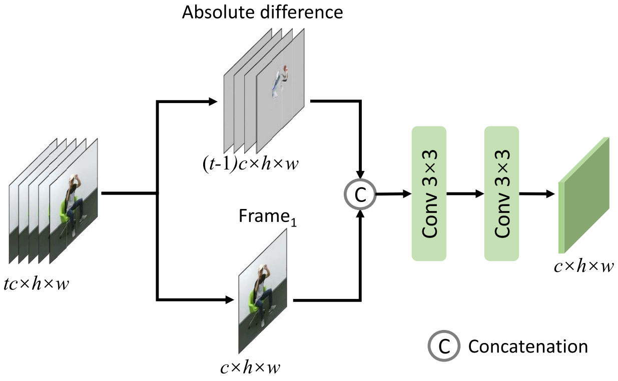

FIGURE 2. Spatial-temporal information aggregation module (STIAM). Shown here is a segment of RGB frames aggregated into one frame size.

advance. The aggregated data is actually the feature map of a shallow network, STIAM is integrated with the backbone network behind, and the entire MMINet is end-to-end trainable. We use TSN [3] and TSM [4] successively as backbone to verify the universality of the approach. Despite embedding STIAM and CMIM, the entire network requires less than 1% extra computation cost based on the backbones, as detailed in subsection C of section V.

B. SPATIAL-TEMPORAL INFORMATION AGGREGATION MODULE

The aggregated RGB and depth data are constructed from segmented RGB-D sequence by STIAM, which aims to utilize motion information in densely sampled RGB-D sequences with minimal computational effort.

As shown in Fig. 2, taking video frames as an example, t frames sampled in a segment are fed into STIAM. Since frame difference has shown effectiveness in presenting temporal information in previous work [3], [7], the absolute difference of adjacent frames is calculated to assist the preliminary features extraction on the time dimension, which also makes the model focus on the motion salient regions. Meanwhile, to retain spatial information, an original frame is also fed into the subsequent steps. Then, instead of constructing hand-crafted spatial-temporal descriptors [7], [14], two convolution layers are employed to compress tc channels into c channels. The convolution operation provides the most effective way to aggregate the information of t adjacent frames. The key calculation can be described as:

$$M = \sum_{i=1}^{tc} C_i * F_i, \tag{1}$$

where M denotes a channel of output features, C is a channel of input, F is the corresponding channel of the filter, and $*$ indicates convolution. The cumulative operation endows M with the information of multiple channels in the time dimension, and the two-layer convolution makes M have sufficient receptive field for the original input in the spatial dimension. Most importantly, compared with the calculation method with fixed manual motion descriptors, the parameters of the convolution layer are constantly optimized with training, so that the most appropriate information can be aggregated. The visualization of aggregation data is shown in Fig. 3, we normalized the aggregated data into RGB image for observation. It can be seen that the aggregation data highlights the region where the movement occurs in a smooth way. Even subtle actions can be accurately captured, such as ‘‘phone call’’.

C. CROSS-MODALITY INTERACTIVE MODULE

To fully exploit the complementary advantages of multimodal, a plug-and-play cross-modality interaction module

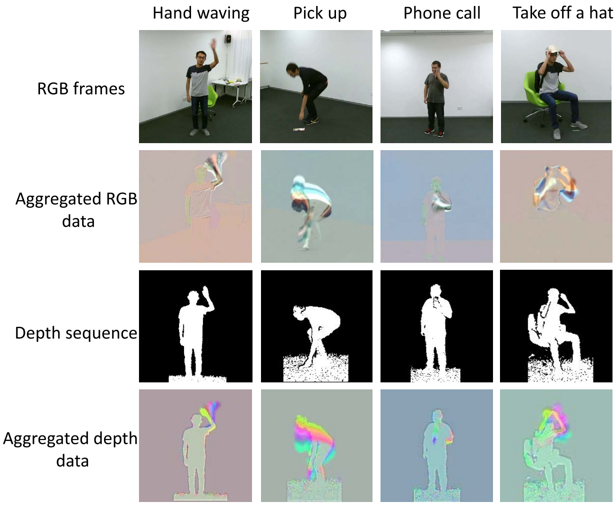

FIGURE 3. Visualization of the aggregated data from STIAM.

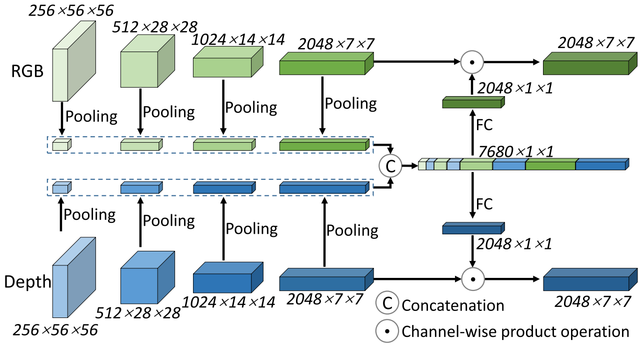

FIGURE 4. Cross-modality interaction module (CMIM).

is designed in the form of neural network. It can be easily used in different feature levels, and even allow the CNN branch of each modality to be initialized with the existing weight. As shown in Fig. 4, CMIM receives the features of each layer of the two CNN branches as input, and learns global multi-modal embedding and uses this embedding to recalibrate the features of the decision layer. Specifically, as suggested by SENet [58], the spatial dimension of the feature maps from each layer are squeezed into 1×1 to describe its channel-level features using global average pooling, the squeeze operation is performed as follows,

$$X = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} M(i,j), \tag{2}$$

where M is a channel of feature map, H and W denote the height and width of M, respectively, X is the result of global average pooling for M. The entire feature map is transformed to a one-dimensional vector. Then, the channel-level representations of different layers are aggregated into a joint representation by a concatenation operation. Finally, two fully-connected layers are utilized to generate excitation signals to recalibrate the channel-wise features of the decision layers of the two modalities, respectively. The CMIM establishes a global fusion mechanism at a light computational cost, which fully combines complementary information from different levels of the two modalities to participate in decision-making.

D. LOSS FUNCTION

To jointly optimite the MMINet, a dual loss function is established based on the standard cross-entropy loss. The $Loss_{RGB}$ and $Loss_{Depth}$ are formulated as

$$Loss_{RGB}(Y, N) = -\sum_{n=1}^{N} Y_c log(Y'_{RGB}), \tag{3}$$

$$Loss_{Depth}(Y, N) = -\sum_{n=1}^{N} Y_c log(Y'_{Depth}), \tag{4}$$

where N is the number of action categories, Y_c is the one hot vector of true label, Y'_{RGB} and Y'_{Depth} are the class probability scores of RGB and Depth streams, respectively. To combine the $Loss_{RGB}$ and $Loss_{Depth}$, the final loss function is denote as

$$Loss(Y, N) = Loss_{RGB} + Loss_{Depth}. \tag{5}$$

IV. EXPERIMENTS

A. DATASETS

To verify the effectiveness and universality of our approach, extensive experiments are conducted using two backbones (TSN and TSM) on three RGB-D action recognition datasets: NTU RGB+D 60 [59], NTU RGB+D 120 [60] and PKU-MMD [61].

The NTU RGB+D 60 dataset consists of 56,880 action samples covering 60 classes, which are performed by 40 distinct subjects and captured by three Microsoft Kinect v2 cameras from different views concurrently. Two evaluation protocols are recommended: cross-subject (C-Sub) and cross-view (C-view) evaluation protocols. For C-sub protocol, action samples performed by 20 subjects are picked for training and the other 20 subjects for testing. For C-view protocol, samples captured by cameras 2 and 3 are picked for training, and captured by camera 1 for testing.

NTU RGB+D 120 contains totaling 114,480 action samples covering 120 classes, which are performed by 106 subjects captured by three Microsoft Kinect v2 cameras from different views concurrently. There are two protocols for evaluating models: cross-subject (C-sub) and cross-setup (C-set). For C-sub protocol, samples performed by 53 subjects are picked for training and the rest 53 subjects for testing. For C-set protocol, the action samples with even setup IDs are picked for training, and action samples with odd setup IDs for testing.

PKU-MMD contains totaling 20,734 action samples covering 51 classes, which are performed by 66 subjects and captured by three Microsoft Kinect v2 cameras from left, middle, and right views concurrently. There are two protocols for evaluating models: cross-subject (C-sub) and cross-view (C-view). For C-sub protocol, samples performed by 57 subjects are picked for training and the rest 9 subjects for testing. For C-view protocol, samples captured by middle and right

TABLE 1. Results of different *t* on NTU RGB+D 60 dataset. Notation for the header: D: Depth.

t	Modality	C-sub	C-view
2	RGB	89.9	94.4
4	RGB	91.4	95.0
6	RGB	**91.9**	**95.4**
8	RGB	91.8	94.9
2	D	88.5	87.9
4	D	89.7	89.6
6	D	**90.6**	90.3
8	D	**90.6**	**90.4**

cameras are picked for training, and the samples captured by left cameras for testing.

B. IMPLEMENTATION DETAILS

The backbone is initialized with ImageNet pre-trained ResNet-50 [62], and the entire model is fine-tuned using SGD with Nesterov momentum (0.9) for 60 epochs. We follow the data augmentation and segment sampling strategy in TSN [3]. The segment k is set to 3. For NTU RGB+D 60 dataset, the initial learning rate is set to 0.01, batchsize to 32. Besides, for NTU RGB+D 120 and PKU-MMD datasets, the initial learning rate is set to 0.001, batchsize to 64. All experiments divide the learning rate by 10 at 25th, 40th and 50th epochs. A dropout layer with regularization ratio set to 0.5 is employed to alleviate overfitting.

C. OPTIMAL SAMPLING DENSITY

The most crucial parameter governing the STIAM is the number of frames sampled from each segment: t. Increasing t is expected to improve the recognition performance of the models. In experiments, we vary the number of t from 2 to 8 and evaluate the recognition performance on TSM + STIAM model. The recognition accuracy on NTU RGB+D 60 dataset are summarized in Table 1. It can be observed that increasing the t generally lead to better performance. When $t = 6$, the performance saturate. Thus, the t is set to 6 in the following experiments.

D. ABLATION EXPERIMENTS

To testify the effectiveness of different modules in our approach, extensive ablation experiments are conduct on three datasets. Specifically, to validate the contribution of each module, we show the performance of each individual module and the combination of all modules (MMINet) in Table 2. Moreover, to demonstrate the universality of our approach, we conduct the experiments on two backbones.

For NTU RGB+D 60 dataset, when using TSN as backbone, STIAM improve the recognition accuracy of RGB and depth modalities on C-sub protocol by 17.3% and 11.7%, and improve the accuracy of RGB and depth modalities on C-view protocol by 14.9% and 13.7%, respectively. When using dual modalities, STIAM delivers 12.1% and 11.8% improvement in accuracy on C-sub and C-view protocols, CMIM delivers 0.7% and 2.7% improvement over C-sub and C-view protocols, respectively. Compared with TSN, MMINet-TSN improves the accuracy on C-sub and C-sub protocols by 12.3% and 13.1% respectively. When using TSM as backbone, STIAM improve the recognition accuracy of RGB and depth modalities on C-sub protocol by 5.6% and 4.4%, and improve the accuracy of RGB and depth modalities on C-view protocol by 3.9% and 5.9%, respectively. When utilizing two modalities, STIAM delivers 3.5% and 3.4% improvement in accuracy on C-sub and C-view protocols, CMIM delivers 0.6% and 1.0% improvement over C-sub and C-view protocols, respectively. Compared with TSM, MMINet-TSM improves the accuracy on C-sub and C-sub protocols by 4.1% and 4.0% respectively.

For NTU RGB+D 120 dataset, our approach also show impressive performance on the NTU RGB+D 120 dataset. Similar to the results on the NTU RGB+D 60 dataset, STIAM significantly improves the recognition accuracy on both single modality and multi-modal, CMIM further improves the performance of the model by enhancing the multi-modal fusion. Compared with TSN, MMINet-TSN improves the recognition accuracy on C-Sub and C-Set protocols by 12.3% and 12.2% respectively. Compared with TSM, MMINet-TSM

TABLE 2. Result of ablation experiment using TSN [3] and TSM [4] as the backbone, respectively. Notation for the header: D: Depth.

Method	Modality	NTU RGB+D 60		NTU RGB+D 120		PKU-MMD	
		C-sub (%)	C-view (%)	C-sub (%)	C-set (%)	C-sub (%)	C-view (%)
TSN [3]	RGB	70.5	75.7	65.1	62.4	82.1	79.6
TSN [3]	D	75.0	70.1	70.1	69.7	82.7	82.6
TSN [3]	RGB + D	78.9	79.9	76.7	77.0	85.0	85.7
TSN + CMIM	RGB + D	79.6	82.6	79.7	80.0	87.2	88.1
TSN + STIAM	RGB	87.8	90.6	84.7	82.2	89.1	87.6
TSN + STIAM	D	86.7	83.8	80.9	81.7	89.4	89.1
TSN + STIAM	RGB + D	91.0	91.7	88.3	88.8	91.4	91.0
MMINet-TSN	RGB + D	**91.2**	**93.0**	**89.0**	**89.2**	**92.6**	**92.1**
TSM [4]	RGB	86.3	91.5	84.1	80.0	90.4	88.9
TSM [4]	D	86.2	84.3	80.8	79.1	89.3	89.3
TSM [4]	RGB + D	90.2	92.5	88.4	87.5	91.7	92.6
TSM + CMIM	RGB + D	90.8	93.5	88.8	90.4	92.0	92.9
TSM + STIAM	RGB	91.9	95.4	86.6	86.3	92.4	92.2
TSM + STIAM	D	90.6	90.2	86.6	86.1	91.5	90.9
TSM + STIAM	RGB + D	93.7	95.9	90.9	91.5	93.1	93.7
MMINet-TSM	RGB + D	**94.3**	**96.5**	**91.7**	**92.6**	**93.6**	**94.2**

TABLE 3. Comparison of the proposed approach and previous works on NTU RGB+D 60 dataset. Notation for the header: D: Depth, S: Skeleton.

Method	Modality	C-sub(%)	C-view(%)
2 Layer P-LSTM [59]	S	62.9	70.3
ST-LSTM+TrustGate [63]	S	69.2	77.7
GCA-LSTM network [64]	S	74.4	82.8
VA-LSTM [65]	S	79.4	87.6
HCN [66]	S	86.5	91.1
Pose-drive Attention [67]	RGB + S	84.8	90.6
Deep Bilinear [45]	RGB + D + S	85.4	90.7
DDI+DDNI+DDMNI [68]	D	87.1	84.2
SSSCA-SSLM [10]	RGB + D	74.9	-
c–ConvNets [46]	RGB + D	86.4	89.1
Glimpse Cloud [69]	RGB + D	86.6	93.2
SC-ConvNets [16]	RGB + D	89.4	91.2
AGC-LSTM [70]	S	89.2	95.0
Separable STA [71]	RGB + S	92.2	94.6
P-I3D [72]	RGB + S	93.0	95.4
VPN [47]	RGB + S	93.5	96.2
TSN [3]	RGB + D	78.9	79.9
MMINet-TSN (ours)	RGB + D	91.2	93.0
TSM [4]	RGB + D	90.2	92.5
MMINet-TSM (ours)	RGB + D	**94.3**	**96.5**

TABLE 4. Comparison of the proposed approach and previous works on NTU RGB+D 120 dataset. Notation for the header: D: Depth, S: Skeleton.

Method	Modality	C-sub(%)	C-set(%)
Internal Feature Fusion [73]	S	58.2	60.9
FSNet [74]	S	59.9	62.4
skeleton visualization [60]	S	60.3	63.2
two-stream attention LSTM [75]	S	61.2	63.3
body pose evolution map [60]	S	64.6	66.9
ResNet-101 [62]	RGB + D	56.5	54.1
Res3D-101 [9]	RGB + S	81.3	83.4
J-ResNet-CMCB [15]	RGB + D	82.8	83.6
Separable STA [71]	RGB + S	83.8	82.5
SC-ConvNets [16]	RGB + D	86.9	87.7
VPN [47]	RGB + S	86.3	87.8
TSN [3]	RGB + D	76.7	77.0
MMINet-TSN (ours)	RGB + D	89.0	89.2
TSM [4]	RGB + D	88.4	87.5
MMINet-TSM (ours)	RGB + D	**91.7**	**92.6**

improves the accuracy of C-sub and C-set protocols by 3.3% and 5.1%, respectively.

For PKU-MMD dataset, STIAM and CMIM consistently improve the recognition performance on two backbones significantly. Compared with TSN, MMINet-TSN improves the recognition accuracy on C-Sub and C-view protocols by 7.6% and 6.4% respectively. Compared with TSM, MMINet-TSM improves the accuracy of C-sub and C-view protocols by 1.9% and 1.6%, respectively.

In summary, STIAM and CMIM modules enhance the capability of the model in extracting spatial-temporal features and learning multi-modal complementary information, which greatly improve the recognition accuracy on both TSN and TSM backbones.

E. COMPARISONS WITH THE STATE-OF-THE-ART APPROACHES

After analyzing the effect of each module in MMINet, we now compare our approach against the state-of-the-art works on three datasets. There are some approaches [78] use the object detection model to crop the person bounding boxes in advance to eliminate most of the background interference. This allowed them to achieve higher recognition accuracy. To be fair, we will only compare with approaches without object detection in this article. The details are shown in Table 3, Table 4 and Table 5.

TABLE 5. Comparison of the proposed approach and previous works on PKU-MMD dataset. Notation for the header: D: Depth, S: Skeleton.

Method	Modality	C-sub	C-view
LSTM [76]	S	83.7	91.0
SA-LSTM [76]	S	86.3	91.4
TA-LSTM [76]	S	86.6	92.3
Bi-LSTM [77]	S	86.5	92.2
STA-LSTM [76]	S	86.9	92.6
J-ResNet-CMCB [15]	RGB + D	91.4	91.4
SC-ConvNets [16]	RGB + D	92.1	93.2
TSN [3]	RGB + D	85.0	85.7
MMINet-TSN (ours)	RGB + D	92.6	92.1
TSM [4]	RGB + D	91.7	92.6
MMINet-TSM (ours)	RGB + D	**93.6**	**94.2**

For NTU RGB+D 60 dataset, as shown in Table 3, our approach achieves 94.3% and 96.5% recognition accuracy, which outperforms other RGB-D based methods by 4.9% and 3.3%, and outperform the state-of-the-art by 0.8% and 0.3% on C-sub and C-view evaluation protocols, respectively. The comparison results on NTU RGB+D 120 dataset are shown in Table 4 It can be seen that our approach achieves 91.7% and 92.6% recognition accuracy, which outperform other methods by 3.3% and 4.8% on C-sub and C-set evaluation protocols, respectively. Table 5 presnets the comparison of out approach with previous works. Our approach consistently show impressive superiority, which achieves 93.6% and 94.2% recognition accuracy, and outperforms the state-of-the-art by 1.5% and 1.0% on C-sub and C-view evaluation protocols, respectively.

V. ANALYSIS AND DISCUSSION

To verify the advantages of our method in recognizing subtle actions, we list the accuracy gain of each action. Then we analyze the specific performance of our approach on different actions through confusion matrix. Finally, we compare the computational cost of our approach with state-of-the-art approaches.

A. ADVANTAGES IN RECOGNIZING SUBTLE ACTIONS

Extensive experimental results on three datasets with different sizes and evaluation protocols have verify the effectiveness of our MMINet. To figure out the strengths of our approach, we list the gain of our MMINet-TSM with respect to the TSM on NTU RGB+D 60 dataset C-sub protocol, NTU RGB+D 120 dataset C-sub protocol and PKU-MMD dataset C-sub protocol, which are shown in Fig. 5, Fig. 6 and Fig. 7 respectively. As shown in Fig. 5, our approach outperforms the baseline for most actions. In addition, the biggest improvement in accuracy are found in several subtle actions, such as "drink water" (8.8%), "clapping" (9.5%), "reading" (10.6%), "writing" (11.8%), and "chest pain" (9.4%). The same phenomenon can also be seen in the NTU RGB+D 120 dataset. As shown in Fig. 6, our approach

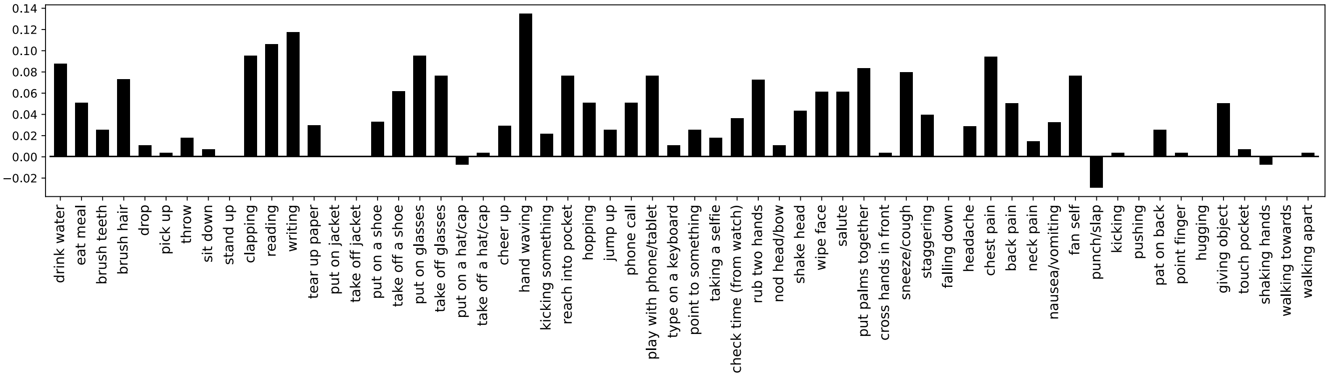

FIGURE 5. Gain on recognition accuracy of MMINet-TSM with respect to TSM on NTU RGB+D 60 dataset C-sub protocol.

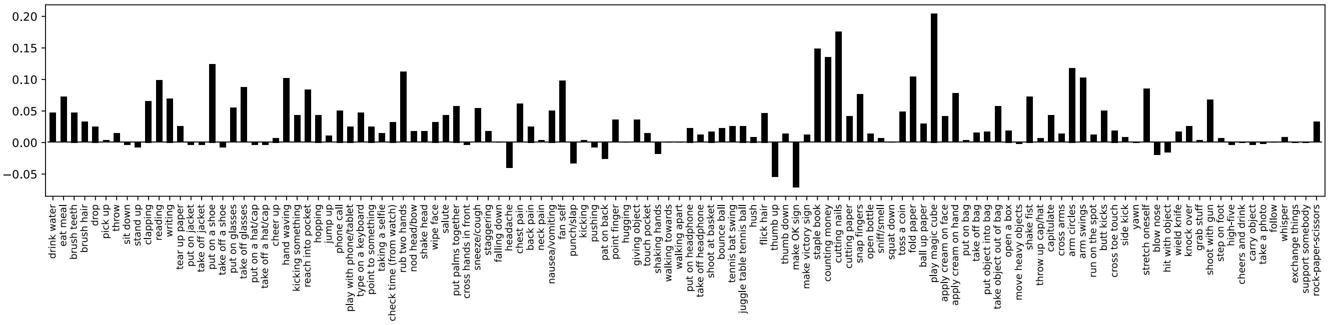

FIGURE 6. Gain on recognition accuracy of MMINet-TSM with respect to TSM on NTU RGB+D 120 dataset C-sub protocol.

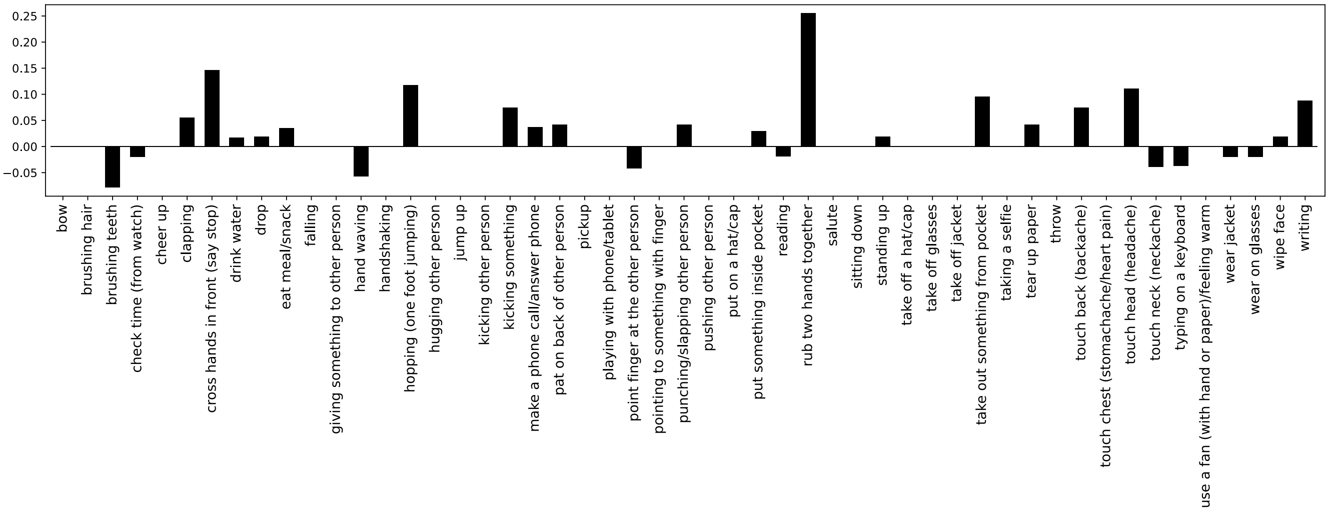

FIGURE 7. Gain on recognition accuracy of MMINet-TSM with respect to TSM on PKU-MMD dataset C-sub protocol.

significantly improves the accuracy of some subtle actions, such as "staple book" (14.9%), "counting money" (13.5%), "cutting nails" (17.6%), and "play magic cube" (20.5%). The accuracy gain results on the PKU-MMD dataset are shown Fig. 7, where the subtle actions also benefit the most, such as "cross hands in front (say stop)" (14.9%), "hopping (one foot jumping)" (11.8%), "rub two hands together" (25.5%), and "touch head (headache)" (11.1%). In conclusion, our method does have unique advantages in recognizing subtle actions. Undesirably, the accuracy of a few actions is worse, such as "punch/slap" (−2.9%) in NTU RGB+D 60 dataset; "headache" (−3.9%), "thumb up" (−5.4%), and "make OK sign" (−7.1%) in NTU RGB+D 120 dataset; "brushing teeth" (−7.8%), "hand waving" (−5.7%), and "point finger at the other person" (−4.1%) in PKU-MMD dataset.

B. ACTIONS THAT ARE EASILY CONFUSED

To find out which actions are prone to being misclassified, we list the confusion matrices under our approach on three

TABLE 6. Comparisons of computation complexity and recognition accuracy with the state-of-the-art approaches. For Separable STA, P-I3D and VPN, we only know the computation cost of RGB streams, that is why the number of FLOPs followed by "+".

Method	Modality	FLOPs(G)	Param.(M)	NTU RGB+D 60		NTU RGB+D 120		PKU-MMD	
				C-sub (%)	C-view (%)	C-sub (%)	C-set (%)	C-sub (%)	C-view (%)
AGC-LSTM [70]	S	54.4	45.7	89.2	95.0	-	-	-	-
MS-G3D [79]	S	99.8	12.8	91.5	96.2	86.9	88.4	-	-
DGNN [80]	S	126.8	8.2	89.9	96.1	-	-	-	-
Separable STA [71]	RGB + S	108 +	12.4 +	92.2	94.6	83.8	82.5	-	-
VPN [47]	RGB + S	108 +	12.4 +	93.5	96.2	86.3	87.8	-	-
P-I3D [72]	RGB + S	324 +	37.1 +	93.0	95.4	-	-	-	-
TSN [3]	RGB + D	24.72	47.3	78.9	79.9	76.7	77.0	85.0	85.7
MMINet-TSN (ours)	RGB + D	24.96 (+0.97%)	76.6	91.2 (+12.3)	93.0 (+13.1)	89.0 (+12.3)	89.2 (+12.2)	92.6 (+7.6)	92.1 (+6.4)
TSM [4]	RGB + D	24.72	47.3	90.2	92.5	88.4	87.5	91.7	92.6
MMINet-TSM (ours)	RGB + D	24.96 (+0.97%)	76.6	94.3 (+4.1)	96.5 (+4.0)	91.7 (+3.3)	92.6 (+5.1)	93.6 (+1.9)	94.2 (+1.6)

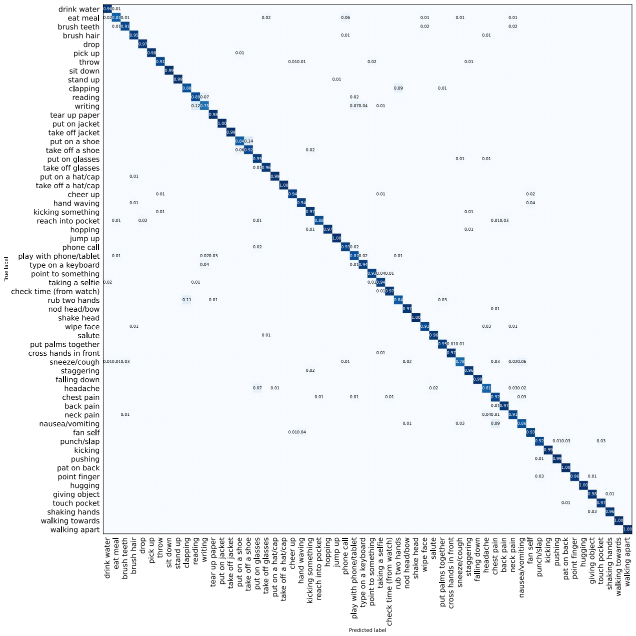

FIGURE 8. Confusion matrix of MMINet-TSM on C-sub proposal of NTU RGB+D 60 dataset.

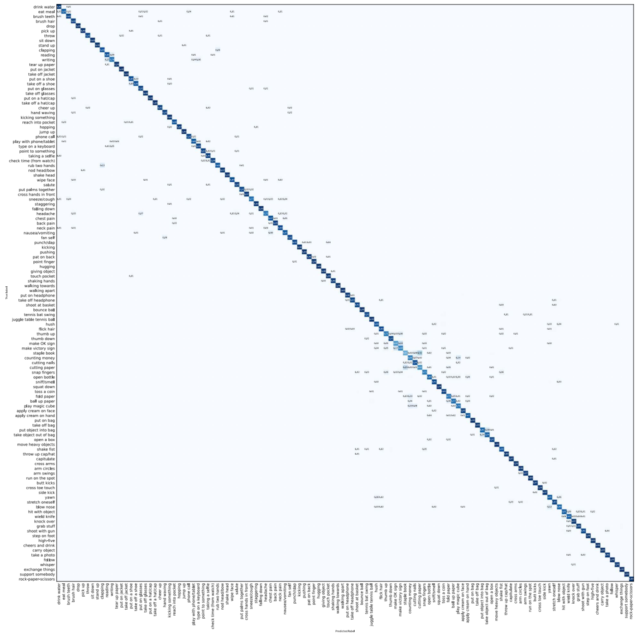

FIGURE 9. Confusion matrix of MMINet-TSM on C-sub proposal of NTU RGB+D 120 dataset.

datasets in Fig. 8, Fig. 10, and Fig. 9. As shown in Fig. 8, most of the actions in the NTU RGB+D 60 dataset with C-Sub protocol can be accurately recognized, and only a few actions are prone to be confused. These confusions tend to occur between actions that have the same body posture with only slight differences at the extremities, such as "writing" and "reading" are easily misclassified as each other, "eat meal" is often misclassified as "phone call", "put on a shoe" and "take off a shoe" are easily misclassified as each other, "sneeze/cough" is easily be misclassified as "nausea/vomiting". Fig. 9 shows the confusion matrix on NTU RGB+D 120 dataset C-Sub protocol. It can be seen that our approach can distinguish most of the actions, but there are still a few actions that easily be confused. These actions have slight distinction only on the fingers, such as "make OK sign" and "make victory sign", or doing the similar operation on similar items, such as "counting money" and "cutting paper". Fig. 10 shows the confusion matrix on PKU-MMD dataset C-Sub protocol. It can be seen that our approach achieves good recognition performance on most actions. The most confusing actions usually have similar appearance and magnitude to each other, such as "clapping" and "rub two hands together" are easily misclassified as each other, "hand waving" is often misclassified as "taking a selfie", "tear up paper" is often misclassified as "typing on a keyboard".

C. COMPUTATION COMPLEXITY OF OUR MODEL

As shown in Table 6, we compare the computational complexity and recognition accuracy of MMINet with several state-of-the-art approaches, including the baselines. Compare to the baselines, MMINet achieves recognition accuracy gains of 1.6% to 13.1% with less than 1% extra computation cost across various evaluation protocols of three datasets. In other words, our approach greatly enhances the spatial-temporal modeling capability and multi-modal information fusion capability of the model with slight computational cost. Since most of the previous methods do not give the computational cost, we can only infer part of the computation cost from the backbone network they use, but even so, the computational cost of MMINet is still much

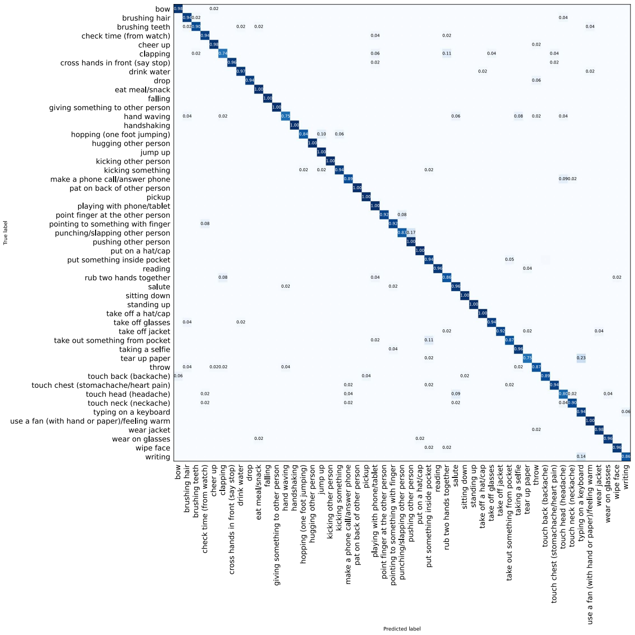

FIGURE 10. Confusion matrix of MMINet-TSM on C-sub proposal of PKU-MMD dataset.

less than that of several state-of-the-art approaches [47], [70], [71], [72], [79], [80].

VI. CONCLUSION

This paper proposes a approach for RGB-D based action recognition. Our approach firstly uses convolutional network to compress the spatial-temporal features of multiple frames to the size of one frame, which avoids the loss of subtle motion information and effectively enhances the recognition for subtle actions. At the same time, we design a cross-modality interaction module, which integrates the features of different levels of the two network branches to recalibrate the channel-wise features, enable the model make full use of the complementary information of different modalities and greatly improve the recognition performance. Extensive experiments on three large multi-modal datasets verify the effectiveness and superiority of our approach. More importantly, our approach provides a general architecture for multi-modal fusion that can be extended to more modalities in the future, such as fusion of RGB and skeleton, fusion of depth and skeleton, etc. The key is to embed the STIAM and CMIM at the appropriate network layers. In addition, other topics in the field of computer vision, such as person re-identification, temporal action detection, can also use multi-modal fusion technology to improve the performance.

REFERENCES

[1] Z. Cai, L. Wang, X. Peng, and Y. Qiao, "Multi-view super vector for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 596–603.

[2] S. Afshar and A. A. Salah, "Facial expression recognition in the wild using improved dense trajectories and Fisher vector encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 1517–1525.

[3] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[4] J. Lin, C. Gan, K. Wang, and S. Han, "TSM: Temporal shift module for efficient and scalable video understanding on edge devices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2760–2774, May 2022.

[5] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 906–915.

[6] T. Liu, J. Kong, and M. Jiang, "RGB-D action recognition using multimodal correlative representation learning model," *IEEE Sensors J.*, vol. 19, no. 5, pp. 1862–1872, Mar. 2019.

[7] P. Khaire, P. Kumar, and J. Imran, "Combining CNN streams of RGB-D and skeletal data for human activity recognition," *Pattern Recognit. Lett.*, vol. 115, pp. 107–116, Nov. 2018.

[8] M. Abavisani, H. R. V. Joze, and V. M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1165–1174.

[9] H. Wu, X. Ma, and Y. Li, "Spatiotemporal multimodal learning with 3D CNNs for video action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1250–1261, Mar. 2022.

[10] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in RGB+D videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1045–1058, May 2018.

[11] C. Liang, D. Liu, L. Qi, and L. Guan, "Multi-modal human action recognition with sub-action exploiting and class-privacy preserved collaborative representation learning," *IEEE Access*, vol. 8, pp. 39920–39933, 2020.

[12] Q. Li, W. Yang, X. Chen, T. Yuan, and Y. Wang, "Temporal segment connection network for action recognition," *IEEE Access*, vol. 8, pp. 179118–179127, 2020.

[13] Z. Gao, H. Zhang, G. P. Xu, and Y. B. Xue, "Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition," *Neurocomputing*, vol. 151, pp. 554–564, Mar. 2015.

[14] B. Fernando, E. Gavves, J. O. M., A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 773–787, Apr. 2016.

[15] J. Cheng, Z. Ren, Q. Zhang, X. Gao, and F. Hao, "Cross-modality compensation convolutional neural networks for RGB-D action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1498–1509, Mar. 2022.

[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, Apr. 2021.

[17] Y. Wan, Z. Yu, Y. Wang, and X. Li, "Action recognition based on two-stream convolutional networks with long-short-term spatiotemporal features," *IEEE Access*, vol. 8, pp. 85284–85293, 2020.

[18] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1932–1939.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[20] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2006, pp. 428–441.

[21] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.

[22] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 9912, 2016, pp. 20–36.

[23] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE Int. Conf. Comput. Vis. (CVPR)*, Oct. 2019, pp. 7083–7093.

[24] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11205. Cham, Switzerland: Springer, Sep. 2018, pp. 831–846.

[25] L. Wang, Z. Tong, B. Ji, and G. Wu, "TDN: Temporal difference networks for efficient action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1895–1904.

[26] X. Ji, J. Cheng, D. Tao, X. Wu, and W. Feng, "The spatial Laplacian and temporal energy pyramid representation for human action recognition using depth sequences," *Knowl.-Based Syst.*, vol. 122, pp. 64–74, Apr. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705117300461

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[28] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[29] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6202–6211.

[30] F. Wang, G. Wang, Y. Huang, and H. Chu, "SAST: Learning semantic action-aware spatial-temporal features for efficient action recognition," *IEEE Access*, vol. 7, pp. 164876–164886, 2019.

[31] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5534–5542.

[32] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6450–6459.

[33] D. Tran, H. Wang, M. Feiszli, and L. Torresani, "Video classification with channel-separated convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5551–5560.

[34] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[35] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2799–2813, Dec. 2017.

[36] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, "A comprehensive study of deep video action recognition," 2020, *arXiv:2012.06567*.

[37] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1390–1399.

[38] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 595–604.

[39] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 568–576.

[40] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4305–4314.

[41] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[42] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4768–4777.

[43] X. Ji, J. Cheng, W. Feng, and D. Tao, "Skeleton embedded motion body partition for human action recognition using depth sequences," *Signal Process.*, vol. 143, pp. 56–68, Feb. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168417302980

[44] R. Zhao, H. Ali, and P. van der Smagt, "Two-stream RNN/CNN for action recognition in 3D videos," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 4260–4267.

[45] J.-F. Hu, W.-S. Zheng, J. Pan, J. Lai, and J. Zhang, "Deep bilinear learning for RGB-D action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 335–351.

[46] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, "Cooperative training of deep aggregation networks for RGB-D action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018, pp. 1–8.

[47] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "VPN: Learning video-pose embedding for activities of daily living," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 72–90.

[48] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10454–10464.

[49] A. Nagrani, C. Sun, D. Ross, R. Sukthankar, C. Schmid, and A. Zisserman, "Speech2Action: Cross-modal supervision for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10317–10326.

[50] S. Song, J. Liu, Y. Li, and Z. Guo, "Modality compensation network: Cross-modal adaptation for action recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 3957–3969, 2020.

[51] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13286–13296.

[52] L. Liu, J. Chen, H. Wu, G. Li, C. Li, and L. Lin, "Cross-modal collaborative representation learning and a large-scale RGBT benchmark for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4823–4833.

[53] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 609–617.

[54] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "EPIC-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5492–5501.

[55] X. Hao, S. Zhao, M. Ye, and J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16383–16392.

[56] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Trans. Image Process.*, vol. 29, pp. 9387–9399, 2020.

[57] Z. Zhao, B. Liu, Q. Chu, Y. Lu, and N. Yu, "Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 3520–3528.

[58] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze- and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[59] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[60] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[61] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," 2017, *arXiv:1703.07475*.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[63] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 816–833.

[64] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3671–3680.

[65] Y. Ji, G. Ye, and H. Cheng, "Interactive body part contrast mining for human interaction recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2014, pp. 1–6.

[66] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 786–792.

[67] F. Baradel, C. Wolf, and J. Mille, "Human activity recognition with pose-driven attention to RGB," in *Proc. 29th Brit. Mach. Vis. Conf. (BMVC)*, Newcastle, U.K., Sep. 2018, pp. 1–14. [Online]. Available: https://hal.inria.fr/hal-01828083

[68] P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Ogunbona, "Depth pooling based large-scale 3-D action recognition with convolutional neural networks," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1051–1061, May 2018.

[69] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 469–478.

[70] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1227–1236.

[71] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, "Toyota smarthome: Real-world activities of daily living," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 833–842.

[72] S. Das, A. Chaudhary, F. Bremond, and M. Thonnat, "Where to focus on for human action recognition?" in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 71–80.

[73] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018.

[74] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, "Skeleton-based online action prediction using scale selection network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1453–1467, Jun. 2020.

[75] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[76] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based LSTM networks for 3D action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3459–3471, Jul. 2018.

[77] P. Elias, J. Sedmidubsky, and P. Zezula, "Understanding the gap between 2D and 3D skeleton-based action recognition," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 192–1923.

[78] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2969–2978.

[79] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 140–149.

[80] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7904–7913.

QIN CHENG received the B.E. degree from the University of Shanghai for Science and Technology, China, in 2016. He is currently pursuing the Ph.D. degree with the Guilin University of Electronic Technology, China. His research interests include computer vision, action recognition, and human–machine interaction.

ZHEN LIU received the B.E. degree from the North University of China, Taiyuan, China, in 2020. He is currently pursuing the dual M.E. degree with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, and the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His current research interests include computer vision and action recognition.

ZILIANG REN received the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2017. He is currently with the School of Science and Technology, Dongguan University of Technology, China, as a Lecturer. His current research interests include machine learning and human action recognition.

JUN CHENG (Member, IEEE) received the B.Eng. and M.Eng. degrees from the University of Science and Technology of China, Hefei, China, in 1999 and 2002, respectively, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2006. He is currently with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, as a Professor, and the Director of the Laboratory for Human Machine Control. His current research interests include computer vision, robotics, machine intelligence, and control.

JIANMING LIU received the B.Eng., B.B.A., and M.Eng. degrees from the University of Science and Technology of China, in 1999 and 2002, respectively, and the Ph.D. degree in information engineering from The Chinese University of Hong Kong, in 2006. He is currently a Professor with the School of Computer and Information Security, Guilin University of Electronic Technology, China. His research interests include wireless sensor networks, optical networks, intelligent control, and the applications of queueing theory. He was also a GCOE Research Fellow of Tohoku University, Sendai, Japan, from December 2007 to June 2009.