

Received February 12, 2020, accepted February 25, 2020, date of publication March 2, 2020, date of current version March 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977856

A Discriminative Deep Model With Feature Fusion and Temporal Attention for Human Action Recognition

JIAHUI YU^{1,2}, HONGWEI GAO¹, WEI YANG¹, YUEQIU JIANG¹, WEIHONG CHIN³,
NAOYUKI KUBOTA³, AND ZHAOJIE JU², (Senior Member, IEEE)

¹School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China

²School of Computing, University of Portsmouth, Portsmouth PO1 3HE, U.K.

³Graduate School of Systems Design, Tokyo Metropolitan University, Tokyo 191-0065, Japan

Corresponding authors: Hongwei Gao (ghw1978@sohu.com) and Zhaojie Ju (zhaojie.ju@port.ac.uk)


This work was supported in part by the LiaoNing Province Higher Education Innovative Talents Program Support Project under Grant LR2019058, in part by the LiaoNing Revitalization Talents Program under Grant XLYC1902095, in part by the Chinese Academy of Sciences (CAS) Interdisciplinary Innovation Team under Grant JCTD-2018-11, in part by the DREAM project of EU FP7-ICT under Grant 611391, and in part by the National Natural Science Foundation of China under Grant 51575412, Grant 51575338, and Grant 51575407.

ABSTRACT Activity recognition which aims to accurately distinguish human actions in complex environments plays a key role in human-robot/computer interaction. However, long-lasting and similar actions will cause poor feature sequence extraction and thus lead to a reduction of the recognition accuracy. We propose a novel discriminative deep model (D3D-LSTM) based on 3D-CNN and LSTM for both single-target and interaction action recognition to improve the spatiotemporal processing performance. Our models have several notable properties: 1) A real-time feature fusion method is used to obtain a more representative feature sequence through composition of local mixtures for enhancing the performance of discriminating similar actions; 2) We introduce an improved attention mechanism that focuses on each frame individually by assigning different weights in real-time; 3) An alternating optimization strategy is proposed for our model to obtain parameters with the best performance. Because the proposed D3D-LSTM model is efficient enough to be used as a detector that recognizes various activities, a Real-set database is collected to evaluate action recognition in complex real-world scenarios. For long-term relations, we update the present memory state via the weight-controlled attention module that enables the memory cell to store better long-term features. The densely connected bimodal modal makes local perceptrons of 3D-Conv motion-aware and stores better short-term features. The proposed D3D-LSTM model has been evaluated through a series of experiments on the Real-set and open-source datasets, i.e. SBU-Kinect and MSR-action-3D. Experimental results show that the proposed D3D-LSTM model achieves new state-of-the-art results, including pushing the average rate of the SBU-Kinect to 92.40% and the average rate of the MSR-action-3D to 95.40%.

INDEX TERMS Human action recognition, RGB-D, attention mode, real-time feature fusion, dataset.

I. INTRODUCTION

Human action recognition has gained more interest in the research community and has become a fundamental task in many applications, such as monitoring security [1], gaming entertainment [2], complex object movements [3], [4], smart indoor security systems [5], video streaming [6]–[8], and healthcare [9]. Generally, RGB data, depth data, skeleton data, and mixed data are used to represent human actions.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Ayoub Khan .

The emergence of sensors/cameras achieves efficient action tracking by providing target trajectory and skeleton joints points. Puwein *et al.* proposed wide baselines-based cameras which can accurately record human pose and estimate human action [10]. Slimani *et al.* proposed an automated recognition model for human interaction activities estimation [11]. Zhao *et al.* proposed the SDG model for human action estimation without sufficient labeled and collected a novel skeleton-based dataset [12]. Many effective real-time human tracking systems based on body parts features were proposed by Jalal *et al.* [13], [14], and they also proposed

a human feature representation and extraction method for depth and skeleton data [15].

Since the CNN has achieved great success in image processing, many CNN-based effective networks have been proposed, such as VGG [16], GoogleNet [17], BN-Inception [18] and ResNet [19]. While there are also many deep networks with great performance for action recognition, for example, LSTM for long-term feature modeling [20]; dual-stream neural networks that can process still images and video frames, including spatial and temporal streams [21]; optical flow networks and these improved methods [22]–[24]. The above studies are all for RGB data. Since the Kinect sensor can easily acquire depth data, efforts on the RGB-D dataset have been widely developed.

In recent years, some promising methods for RGB-D-based human action recognition have emerged. Wu *et al.* designed a deep dynamic neural network (DDNN) to implement gesture recognition for multimodal input data. This network can extract spatiotemporal features from depth images [25]. Wang *et al.* proposed a scene flow dynamic model to extract features from RGB-D images by using the ConvNets network [26]. Kim *et al.* proposed a circulatory neural network (PRNN) based on privileged information for deep sequences recognition [27]. Wang *et al.* adapted the DMM to a pseudo-RGB image which converted its spatiotemporal data into texture information, and the model trained by merging three independent ConvNets. They also extracted features in depth image sequences by constructing three different dynamic depth images, namely dynamic depth images, dynamic depth conventional images, and dynamic depth motion conventional images [28], [29]. Rahmani *et al.* proposed a model with infinite sequence learning view-invariant. Each depth image was input into a specific CNN to learn advanced features, and then the action data was transmitted to the model for training [30]. However, these studies have three main limitations: 1) Complex actions recognition is a challenge, such as the combination of several simple actions and long-lasting actions; 2) Poor performance in distinguishing similar actions, such as gestures and timing similar actions; 3) Existing public datasets are not complex enough to represent actual situations.

In this paper, we propose a discriminative deep model (D3D-LSTM) for RGB-D based human gesture recognition to overcome problems that previously mentioned. Our model is almost immune to illumination and occlusion which achieves significant performance in different complex environments. In particular, the D3D-LSTM model achieves a high recognition rate for a variety of RGB-D datasets. The main contributions are summarized as follows.

- 1) A real-time feature fusion method is proposed by combining RGB and depth features more effectively without losing important features. The method achieves better performance in feature fusion which improves recognition accuracy.

- 2) The proposed D3D-LSTM model can deal with the long-term and spatial features of actions more effectively, especially for complex and combined actions, and similar actions recognition.
- 3) The attention mechanism is further improved by assigning a corresponding weight to each element in the feature vector to represent the importance of the element. Each weight is determined by the combination of the upper layer and the current state. This approach improves the recognition rate of long-term complex actions.
- 4) A new RGB-D dataset for action recognition, termed as Real-set, is designed and collected. It is more complex than the current available datasets. Data in Real-set contains changes in illumination intensity, angle, and occlusion, which is more closer to the actual situation.

The remainder of this paper is organized as follows. Section II briefly reviews related work about action recognition methods. Section III introduces the collection process of the RGB-D action dataset. Section IV describes the proposed D3D-LSTM model in detail. Section V reports the experimental sets and results analysis. Section VI concludes the paper and gives the future work.

II. RELATED WORK

A. CONVENTIONAL ACTION RECOGNITION

An ideal recognition model relies on effective learning of action features. In recent years, many different techniques are applied to extract and represent both short-term and long-term spatiotemporal features [31]. For human interaction recognition, the Harris corner points and the histogram [32], and a compact and discriminative video encoding method [33] were proposed to extract features. For objects tracking, chain coding mechanism and centroids point extraction were extended to label body parts [9], [34]. Besides, Kamal *et al.* introduced the hidden Markov model (M-HMM) to fuse spatial depth shape features and temporal joints features for feature representation [35].

CNNs and RNNs have been extensively applied in feature learning. In [16], [17] and [19], the application of CNN to process images achieved great success. Ji *et al.* extended CNN and proposed 3D-CNN [36] that enables CNN to deal with time information. D. Tran *et al.* [37] proposed the optimal convolution kernel size of 3D-CNN, applied to the C3D network of the large-scale datasets, improved the residual network of 3D-CNN, and proposed a Res3D network superior to C3D. Chen *et al.* proposed a lightweight multi-fibre network that optimized network performance [38]. Yang *et al.* proposed an asymmetric 3D-CNN network that combined optical flow frames and RGB features to improve network performance [39]. While Hussein *et al.* improved the 3D-CNN to extract temporal features of action more efficiently [40].

Owing to the poor performance of the CNN extraction temporal features, researchers introduced the LSTM to improve the performance of processing complex actions.

Wang *et al.* proposed a dynamic model for target tracking, which combined the attention mechanism with CNN and LSTM to improve the recognition rate [41]. Ullah *et al.* proposed a two-way LSTM combined with CNN to recognize long-term actions, mainly used sequence stacking in forward and backward propagation, which deepened the ability to understand action intentions [42].

B. SKELETON BASED ACTION RECOGNITION

The skeleton-based method mainly acquires 3D dynamic features, including line features, surface features, and body features. Yang *et al.* determined the classes of actions by calculating the position of 3D joint points, which can effectively recognize static and dynamic actions, used the PCA algorithm for dimensionality reduction, and applied the NBNN algorithm for classification [43]. In the description of static action features, the time domain pyramid covariance descriptor, the sparse coding and the pyramid histogram method were all effective [44]–[46]. Surface-features-based method was mostly inspired by Tang *et al.* They used 3D normal vectors to construct 2D histograms to describe the shape of the target [47]. Recently, Yang *et al.* proposed an adaptive space-time pyramid to divide the depth image and then used the aggregate hypersurface to describe the actions [48].

C. DENSE CONNECTION

DenseNets has been proposed by Corn Huang *et al.* in 2017, which was inspired by the ideas of highway networks and ResNet [19], [49], [50]. The shortcut connection method is a cross-layer connection rather than a sequential connection which is used in all three networks. The purpose is to solve the problem of existing gradient divergence. The shortcuts used in DenseNets are the most efficient. Usually, every 2 or 3 layers would be directly connected using the shortcut. This method can transfer information from the shallow part of the network to the deep part. In particular, partially dense connections can avoid some problems, such as oversized models, excessive parameters, and poor training efficiency, it is shown as Fig.1.

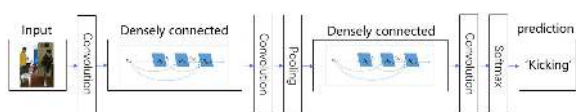


FIGURE 1. Shortcut connection method.The two blue areas use dense connections independently to share data. Generally, the two areas are connected by pooling layers and convolution layers.

In this paper, we extend the traditional DenseNets for static image recognition to 3D-CNN, which can extract the spatiotemporal features of actions. The operation is to extend the convolution kernel from $d \times d$ to $k \times d \times d$. Using the DenseNets connection method, the output characteristics of each layer can be reused, which significantly reduces the amount of calculation. Although the densely connected 3D-CNN is more complex than the C3D, it can improve the

recognition accuracy, and it is much simpler than the method base on ResNet.

D. ATTENTION MODE

Attention Mode provides an effective idea for natural language, image recognition, and big data mining, which is a more popular idea recently [51], [52]. Attention Mode draws on the human visual system, it can quickly scan, lock targets and focus on global images [53]. In action recognition study, the idea of attention mode is to aggregate multiple feature vectors to obtain aggregated features (h). For the first time, Sharma *et al.* introduced attention mechanism into human action recognition, where they focused on the body, clothes and backpacks [54]. Bahdanau *et al.* proposed attention mechanism in the time dimension, and used a weighted summation method [51]. Li *et al.* proposed an end-to-end sequence method for action recognition in video [23]. However, these methods cannot be extended to other studies, and how to comply with the novel model is not addressed.

In the previous studies, the weighted average method was used to obtain the feature weight α_i . Although this method focused on global features, it did not allow the model to know the key element in the feature, and the performance was rarely improved. Therefore, we assign each element a corresponding weight to represent the importance of that element in the feature, which is shown as

$$x_i = \sum_{i=1}^{k \times k} \frac{\exp(W_i h_{t-1})}{\sum_{j=1}^{k \times k} \exp(W_j h_{t-1})} X_{t,i}. \quad (1)$$

III. REAL-SET

In this section, we describe our RGB-D human action dataset which is named Real-set. This dataset is collected to train the proposed model in a real environment. By analyzing the public datasets, we find that although these datasets are large in scale, they ignore the interference that exists in the real world, and the data preprocessing is not sufficient. In this paper, we design and collect a dataset, that is, Real-set, to deal with these problems.

A. DATASET COLLECTION

We collect the RGB-D dataset by applying Kinect 2 which is an RGB-D sensor from Microsoft, and Figure 2 shows the dataset collection process. This sensor can simultaneously capture the RGB data and depth data. Depth data can detect the target from the complex background and be not affected by lighting conditions. The acquisition speed is 30 *fps/S*, the resolution of the image is 640×480 , and the sensor’s acquisition range is 0.8 *m* to 3.5 *m*. All the actions were finished by five volunteers (3 males and 2 females), each action performed at different light intensities, different angles, different backgrounds, and partial occlusions. The dataset combined single-person actions and interaction actions, including horizontal-waving, high-swinging, beating, punching, approaching, kicking, hugs, and shaking hands. Each action was collected with 4000 samples, of which

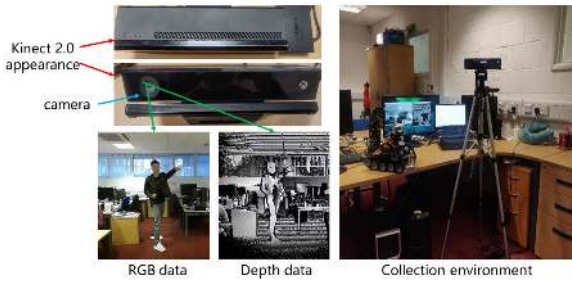


FIGURE 2. Real-set collection. In the Kinect 2.0 sensor, a colour camera, a depth camera (based on the infrared acquisition method), and a microphone are built-in. In theory, the two cameras are acquired simultaneously, and the pixels in the colour image and the depth image correspond one-to-one.



FIGURE 3. Real-set. The first line is RGB data, and the second line is depth data for the corresponding behavior. Each action is performed by different participants under changing conditions, with the goal of simulating a real scene.

3000 were used as training samples and 1000 were used for testing. Therefore, the Real-set has a total of 64,000 video clips, namely, $4000(\text{samples}) \times 8(\text{classes}) \times 2(\text{modes}) = 64000$, and each sample contains both RGB and depth modes. Figure 3 shows some samples in the Real-set.

Compared to the existing datasets, the Real-set has three advantages: 1) More interaction samples: the Real-set has 4000 samples, which is many times than that of other datasets, such as MSR-action-3D dataset (567 samples), SBU-Kinect dataset (300 samples), NTU RGB-D dataset (880 samples); 2) More complex: actors have different body shapes and skin tones, and the light, angle, and scale in the data collection scene are constantly changing and occluded; 3) More effective data: the Real-set has raw data and pre-processed data, which provides convenience for using data in pixel level to get better results and researching data processing methods. The Real-set will be a benchmark dataset for human action recognition based on multi-modal data. The Real-set and trained models will be made available to the public when we finish the skeleton data collection.

B. DATA PREPROCESSING

Most of the public datasets are raw data without preprocessing, which will lead to three problems: both modes of acquisitions are not synchronized; the depth data is partially missing; and the colour data is noisy. Therefore, we have taken three methods to solve the above problems, the data processing flow is shown in Fig. 4.

The sensor is calibrated before acquisition such that the three-dimensional coordinates of RGB data and depth data are in one-to-one correspondence. For $H_d = [X_d, Y_d, Z_d]^T$

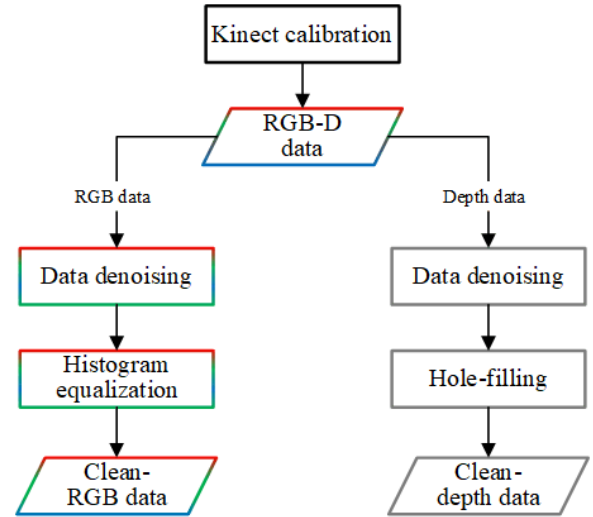


FIGURE 4. Data processing overall flow. After the Kinect is calibrated, RGB-D data is collected. Next, RGB data and Depth data are processed separately: 1) RGB data: after the data is denoised, increasing the fullness of RGB data by applying histogram equalization; 2) Depth data: after the data is denoised, filling depth data holes by applying joint bilateral filtering. Finally, clean-RGB-D data is obtained for model training.

is the coordinates of the depth image, $H_R = [X_R, Y_R, Z_R]^T$ is that of RGB image, the relationship between them is $H_R = PH_d + t$, P is the rotation transformation matrix, and t is the translation vector. By performing a homogeneous transformation on both, Equations (2) and (3) can be obtained. After experimental measurements, the parameters in the two equations are obtained, including $\alpha = 528.32$, $\beta = 527.03$, $i_0 = 320.10$, $j_0 = 257.57$, $t = [25, 2, -2]^T$, and $p = [0.05, -0.01, 0.02]^T$.

$$z \begin{bmatrix} i \\ j \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha & \mu & i_0 & 0 \\ 0 & \beta & j_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (2)$$

$$\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} P & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_d \\ Y_d \\ Z_d \\ 1 \end{bmatrix} \quad (3)$$

When the target depth data is recorded by the infrared camera of the Kinect, the problem of partial area deletion in the depth data is caused by object reflection and diffraction. In this paper, joint bilateral filtering is used to denoise and fill the image to preserve more edge data. The method is as shown in (4), where $C(x, y)$ represents depth data, $P(x, y)$ is a pixel point, $W(x, y)$ is a weight, and G_σ is a Gaussian function. The change of the histogram of the depth data after preprocessing is shown in Fig. 5.

$$B(C(x, y)) = \frac{1}{W_{x,y}} \sum_{(x',y') \in R^{D(x',y')}} G_{\sigma_x}(\|P_{x,y} - P_{x',y'}\|) \times G_{\sigma_y}(\|D_{x,y} - D_{x',y'}\|) \quad (4)$$

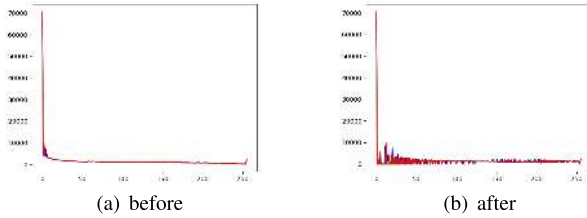


FIGURE 5. Depth data preprocessing. Use additional colour data to fill holes in-depth images. (A) The depth image is empty, some data are missing, and the curve is close to the X-axis. (B) After filled, the data are rich, and the curve is increased.

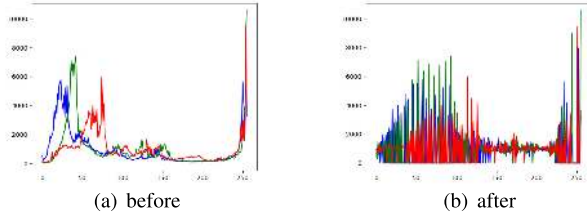


FIGURE 6. RGB data processing. It enhances the contrast between individual pixels in an image. (a) The contrast is weak and the colour is poor in each pixel. (b) The contrast is enhanced, and the colour is rich in each pixel.

The RGB data is processed in a balanced manner. The histogram data of the R/G/B three colour channels are respectively counted, and then the equalization operation is performed. The original channel data is replaced by the mapping of the three channels. The method is as shown in (5) and the effect of preprocessing is shown in Fig. 6.

$$R(x) \rightarrow \hat{R}(x), G(x) \rightarrow \hat{G}(x), B(x) \rightarrow \hat{B}(x) \quad (5)$$

IV. PROPOSED APPROACH

The traditional 3D-CNN can only perform local short-time feature extraction. The traditional LSTM network does not perform well for global long-term feature extraction. To solve these problems, we propose the D3D-LSTM model based on 3D-CNN and LSTM, which introduces the improved attention mechanism and feature fusion method. In this section, we introduce the proposed D3D-LSTM model in detail.

A. NOVEL MODEL STRUCTURE

To better extract the temporal and spatial features of human action, we design the D3D-LSTM model based on 3D-CNN and LSTM, the pipeline of our proposed method is shown as Fig.7. The model consists of three steps, that is, spatiotemporal feature extraction based on 3D-CNN, key temporal feature extraction based on LSTM, and classification.

The huge scale of 3D-CNN often leads to problems such as inefficient training of models and incorrect use of parameters. Therefore, we introduce the idea of dense connection, which allows parameters to be shared in 3D-CNN and improves operation efficiency. A real-time fusion method is adopted to synchronously extract RGB and depth features in 3D-CNN. Immediately after each extraction, it is merged into elements

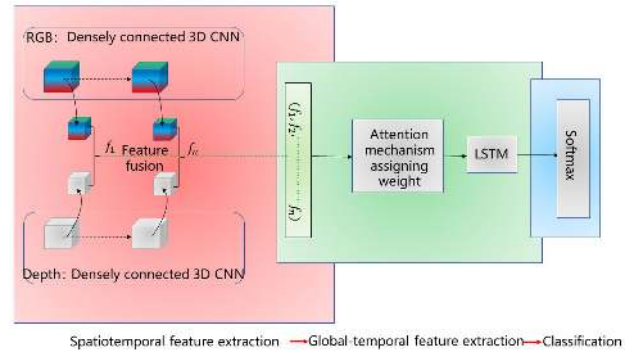


FIGURE 7. The overview of the proposed D3D-LSTM framework. The red area represents spatial features and local short-time feature extraction. The dual-modal features are merged into a one-dimensional feature vector in real-time. The green area represents that each element in the one-dimensional vector is assigned a corresponding weight by attention mechanism to distinguish the importance of each element so that the LSTM can finish focused learning. The blue area represents the final classification and outputs a probability vector.

in the feature vector. Experiments show that this feature extraction method of real-time fusion can obtain more representative feature vectors. Besides, an attention mechanism is introduced in the LSTM. Each element in the fused feature vector is assigned a corresponding weight and then input into the LSTM for training. Finally, the action classification is finished using the classic softmax classifier.

B. DENSELY CONNECTED 3D-CNN

In 2016, Anguelov *et al.* proposed using a convolutional neural network of the same structure to train multiple modal data separately, and feature fusion at appropriate locations to obtain more distinguishing features and enhanced feature robustness [55]. Based on this idea, we design a dual-mode 3D-CNN to train RGB and depth data respectively, and the structure of 3D-CNN in both modes is the same. Taking the process of extracting RGB features as an example, the method is described in detail.

The network consists of five 3D convolution (3D-Conv) layers, two Max-pooling layers, five BN layers, and three dense connection operations. This structure is shown in Fig.8. We introduce the dense connection to the double-module 3D CNN model for feature extraction and fusion. For feature extraction (RGB/Depth), it can speed up model training and feature transfer, and avoid vanishing gradients. For feature fusion, it can provide efficiency features for the fusion process because the previous outputs can affect the following layers. To speed up the convergence of the network and prevent gradient explosion, we add a large number of BN layers, which improves the training efficiency of D3D-LSTM networks.

The pipeline of the network is as follows: 1) Using 64 3D-Conv kernels to extract the features of the input data, and obtain 64 feature maps, and the size of each feature map does not change, that is, $64@32 \times 112 \times 112$; 2) Adopting a $1 \times 2 \times 2$ size kernel to reduce the size of the feature and

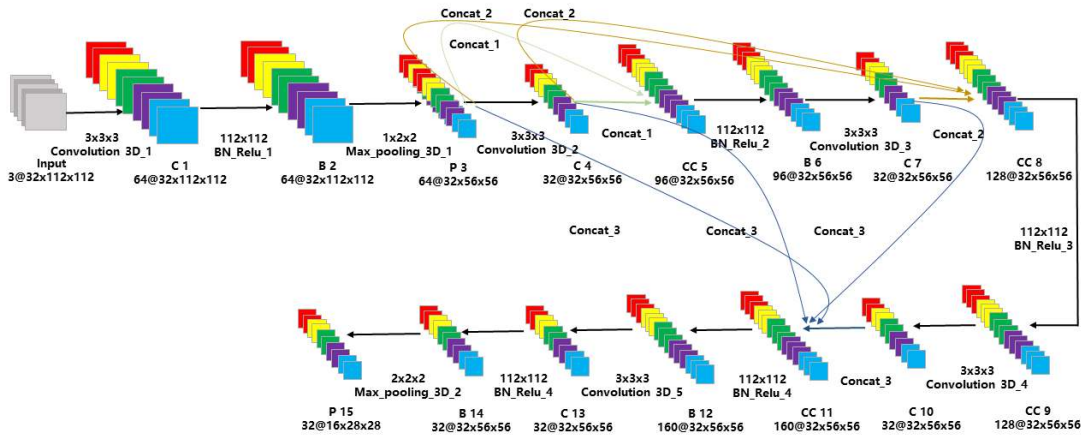


FIGURE 8. Densely connected 3D-CNN network. Each coloured cuboid represents the processed data, and the change in the size of the cuboid represents the change in the size of the data. The yellow, green, and blue connection lines represent three dense connections, respectively. The 3D-Conv kernel size and pooling layer size are obtained through comparative experiments.

the number of parameters, the time dimension is unchanged, and the output feature map size is $64@32 \times 56 \times 56$; 3) The input feature maps are processed by using 32 3D-Conv kernels; 4) Splicing the feature maps in step 2 and 3 to obtain features of 96 channels, their size is unchanged, this operation increases the feature diversity, and reduces the number of the parameters; 5) The third 3D-Conv operation, using 32 convolution kernels, obtains the feature maps with the size of $32@32 \times 56 \times 56$; 6) Splicing the step 2/3/5 into feature maps with the size of $128@32 \times 56 \times 56$; 7) The fourth 3D-Conv operation for extracting 32 key feature maps from 128 sets; 8) The last layer splicing operation, including the features in step 2/3/5/7; 9) The last 3D-Conv operation for extracting 32 feature maps; 10) Selecting a pooled check feature size of $2 \times 2 \times 2$ to reduce the dimension, and outputting the feature maps with the size of $32@16 \times 28 \times 28$.

C. FEATURE EXTRACTION FOR REAL-TIME FUSION

Real-time fusion is more effective than the methods of early-fusion and post-fusion, and the robustness and discrimination of the fused features can be enhanced. This is because the commonality of the RGB and depth features is extracted.

The real-time fusion feature extraction framework is shown in Fig.9. All the large cubes (color/gray) indicate data transition paths with 3D-Convs or max-pooling in the densely connected 3D CNN, and small cubes (color/gray) indicate data fusion operation. After the transition, the output feature maps f_{RGB}^i and f_D^i are fused into a new feature f_i in real-time, while f_{RGB}^i and f_D^i are also input to the next layer, and so on. Then, the new features obtained by each fusion are concatenated into a more representative feature vector (f_1, f_2, \dots, f_n) . Next, the vector is processed by attention-based LSTM to extract global-temporal features, as shown in Figure 6 (green area). Where f_{RGB}^i represents the RGB feature extracted for the i times in the model, and f_D^i is the same. Based on the differences and commonalities between RGB and depth

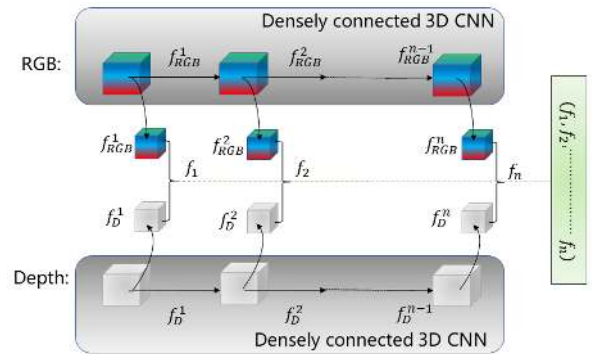


FIGURE 9. Real-time fusion feature extraction network. The gray area represents the dual-modal feature extraction, and the structure is the same. Each cube (color and gray) represents the feature extracted in each frame. Every two corresponding cubes are fused into one element in the feature vector.

features, the following equations are obtained: RGB feature is disassembled into $f_R = f_{1R} + f_{CM}$, depth feature is split into $f_D = f_{1D} + f_{CM}$, and f_{1R} and f_{1D} , that are differences. The sexual part, f_{CM} is the part with the sameness. Therefore, the real-time feature fusion can be expressed as $f = f_{1R} + f_{1D} + f_{CM}$. To complete the recognition task, it is also necessary to obtain the real label L_{true} of the action, which can be obtained by the regression coefficient matrix. The weights corresponding for f_{1R} , f_{1D} , and f_{CM} are W^{1R} , W^{1D} , and W^{CM} , the L_{true} is shown as

$$\left\langle W^{(CM)T} \parallel W^{(1R)T} \parallel W^{(1D)T} \right\rangle^T \times \langle f_{CM} \parallel f_{1R} \parallel f_{1D} \rangle^T = L_{true}. \tag{6}$$

D. ATTENTION MECHANISM BASED LSTM

The attention mechanism allowed the model to focus on the integrity of the input and improve the performance of the model [56]. Based on this idea, we configure the corresponding weights for each input frame. This method is to configure

a larger weight for the frame containing a large amount of action information, so that these frames get the attention of the model, thereby improving the recognition rate.

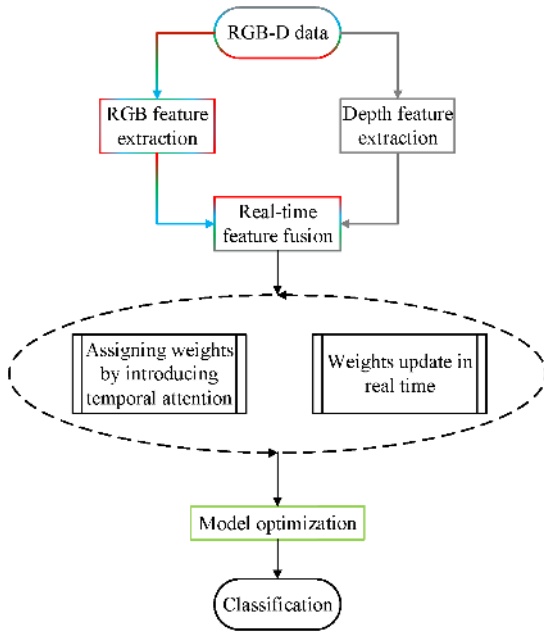


FIGURE 10. Attention mode based flowchart. First, the spatial and local short-term features of the behaviour are extracted for real-time fusion. Then, the attention mode is used to process the fused feature vectors, so that the LSTM has a focused learning process, and the process is iterated in real-time. Finally, model optimization and output classification results.

The pipeline of the LSTM based on the attention mechanism is shown in Fig.10. First, the fusion feature f_i is extracted by 3D-CNN, which is formed by the real-time fusion of RGB and depth features. The fused features are then processed with an improved attention mechanism, assigning corresponding weights to each frame, as shown in (7). Where α_i^t is the weight of the i th element in the fused feature, and the weighted sum of all elements is 1. Therefore, α_i^t can represent the importance of each element. The larger the value, the more critical it is. Finally, how to choose α_i^t is the key to the model.

$$x_i(f) = \sum_{i=1}^N \alpha_i^t f_i \quad (7)$$

In the LSTM, the key is that the last input can affect the next output, and the loop operation can focus on global long-term information. Based on this idea, the choice of α_i^t is also related to the output of the last neuron. The activation function tanh is selected, and the parameters are normalized to obtain the expression of α_i^t , as shown in (8). Where h_{t-1} is the output of the last neuron, and f_i is the dual-stream fusion feature. $h_a, w_a,$ and u_a are the weight matrices obtained during model training, where $w_a \in \mathbb{R}^{n \times n}, u_a \in \mathbb{R}^{n \times 2n}, h_a \in \mathbb{R}^{n \times n}$.

$$\alpha_i^t = \frac{\exp\{\tanh(h_a, w_a, u_a) + f_i h_{t-1}\}}{\sum_{i=1}^N \exp\{\tanh(h_a, w_a, u_a) + f_i h_{t-1}\}} \quad (8)$$

These weights are introduced into the input vector so that the network can focus on each element in the input sequence by its usefulness. α_i^t is also a kind of dynamic weight, which is determined by the output h_{t-1} of the previous moment and the input f_i of the current state, which can more closely represent the importance of each element. In summary, the LSTM structure based on the improved attention mechanism is obtained, as shown in Fig.11. After the global long-time feature is extracted, the vector is input into the softmax classifier for classification, and a probability vector is obtained. Based on the above improvements, the proposed model is more effective for long-term feature processing and improves the recognition rate.

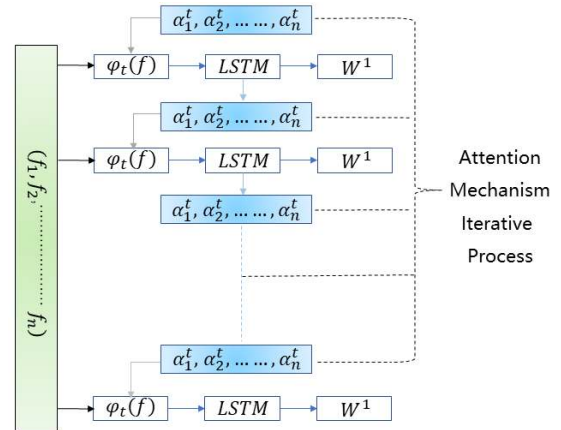


FIGURE 11. Attention mechanism based LSTM network. The blue area represents the weights that have been iteratively updated by the attention model and applied to the input vector. W^1 is the weight generated after each iteration and is used for the final classification.

E. MODEL OPTIMIZATION

The proposed D3D-LSTM model is decomposed into two parts. The 3D-CNN is first optimized and then optimized for the LSTM. Where the former is responsible for extracting and real-time merging RGB and depth features, the latter mainly extracting global long-time features. Comparing with other optimization models, such as SGD [57] and ADAM [14], our model applies the alternating optimization strategy, which uses different methods to update the parameters, including the multi-modal feature adaptive weight learning method and SGD method. The dimension of both features is different in each fusion hierarchy, which would lead to the failure of the multi-feature fusion. We utilized the mapping module method to optimize depth and RGB features. First, the original feature inputs are mapped into the same feature space. Then, each new feature $f_i = [f_{RGB}^i; f_D^i]$ is obtained by applying our feature fusion strategy. Finally, all the f_i are listed in temporal order to be the multi-feature fusion sequence.

The method of optimizing the objective function is adopted to optimize this model. Let the objective function be $min_S = S_R + S_D + S_F$. Where S_R is the cost function of RGB mode, S_D is the cost function of depth mode, and S_F is

the cost function after fusion. When the cost function is minimized, the 3D-CNN model can output the optimal result. (S_R, S_D, S_F) are expressed as (9), (10), and (11), respectively. Where both μ_1 and μ_2 are the weights in the modal and satisfies $\mu_1 + \mu_2 = 1$, W is the transformation matrix of the independent correlation features, f is the characteristic of each modal extraction, α is the weight coefficient of the penalty function, $g(\cdot)$ is the penalty function, λ represents the relationship between feature constraints and supervision in the fusion feature, and L_{true} is true label. During the model optimization, the update of the parameters take place until the model converges, the loop iteration steps are summarized as follows. First, initializing with random values for one part of parameters, including W , W_1 , W_2 , f , while μ_1 and μ_2 are initialized as 0.5. Next, the weights μ_1 and μ_2 are updated in each module (other weights are fixed) by constructing the Lagrangian objective function. In the following iteration process, μ_1 and μ_2 are different because both play different roles in each feature extraction module. Then, the values W_i and f are updated (other weights are fixed) through the gradient descent method.

$$S_R = \mu_1(\|W_1X_1 - f_R\|_F^2 + \|W_1^T f_R - X_1\|_F^2 + \alpha_1 g(f_R)) \quad (9)$$

$$S_D = \mu_2(\|W_2X_2 - f_D\|_F^2 + \|W_2^T f_D - X_2\|_F^2 + \alpha_1 g(f_D)) \quad (10)$$

$$S_F = \lambda(\|W^T f - L_{true}\|_F^2 + \alpha_2 g(\|W\|_{2,1})) \quad (11)$$

Each norm can optimize the objective function. Both (9) and (10) contain two norms. The first norm represents the similarity of the feature and the transformation matrix of the added modal extraction, and the second norm represents the ability of the improved feature to reconstruct the original feature in the reverse direction. Equation (11) takes advantage of the ability to monitor the characteristics of information fusion. Besides, although S_R and S_D seem to be optimized independently of each other in the objective function, f_R and f_D in the two equations contain the same features f_{CM} .

V. EXPERIMENTS AND ANALYSIS

In this section, we conduct numerous experiments and evaluate the proposed D3D-LSTM model on two tasks, that is, similar activity differentiation and complex action recognition. First, two sets of experiments were performed using the Real-set, that is, many conventional algorithms and state-of-the-art methods are tested. Next, used the public datasets to test the proposed model, and compared the experimental results with other advanced results, and comprehensively analyzed the performance of the model. These experiments verify the effectiveness and advancement of our model.

A. EXPERIMENTAL SETTINGS

We use the Real-set, SBU-Kinect and MSR-action-3D data sets, each of which is split into two parts, 70% for training model and 30% for testing model. For fair comparisons,

we operate the *Cross Subject Test*, which can verify results accuracy and generalization. We follow the formulation technique of [58], which includes three concise settings for training and two effective settings for testing. The selected samples for *Cross Subject Test* are applied to the same sampling strategy as [59].

In this model, the two modes that process RGB and depth data use the same structure of the network, sharing the same initialization weights. When training with the Real-set, the initial learning rate is set to 0.001, the learning rate attenuation factor is 0.1/5000 times, the network training step number is 30,000 steps, and the batch value is 16. Since the framework used is the Tensorflow 1.1.4 GPU, the NVIDIA GTX1080 graphics card is mainly used for training. The training epoch is 20, and other parameter settings are empirically obtained $q = 1.5$, $\lambda = 1500$, $\alpha_1 = 2$, $\alpha_2 = 10$, $\gamma = 0.001$.

B. HUMAN ACTION RECOGNITION: REAL-SET

This set of experiments intends to prove the correctness of the proposed model. In particular, the new model combines the traditional networks of 3D-CNN and LSTM and introduces advanced ideas such as attention mechanism and real-time feature fusion. This combined idea has rarely been studied before.

TABLE 1. Verification results on the real-set.

Method	Average rate
RGB-net	88.90%
Depth-net	89.40%
3D-CNN-net	92.70%
New 3D-CNN-net	93.90%
LSTM-net	95.20%
D3D-LSTM	96.40%

In this paper, the D3D-LSTM model is proposed based on some traditional algorithms, using Real-set to test these algorithms. We conducted the following six sets of experiments: 1) Testing the proposed model using the RGB data in the Real-set, namely, RGB-net; 2) Testing the proposed model using the depth data in the Real-set, namely, Depth-net; 3) Testing 3D CNN without real-time feature fusion part using the Real-set, namely, 3D CNN-net. 4) Testing 3D CNN using the Real-set, namely, New 3D CNN-net; 5) Testing the proposed D3D-LSTM model without the attention mechanism using the Real-set, namely, LSTM-net; 6) Testing the proposed D3D-LSTM model using the Real-set. The results of the six groups are shown in Table 1.

When only a single-modal dataset is input, the recognition rate based on the depth dataset is higher than that of the RGB dataset, which proves that the depth data can improve the performance of the action recognition. Comparing the results of the groups 4 and 5, it can be concluded that the method of real-time feature fusion is effective, and its performance is superior to the early-fusion or post-fusion scheme in the traditional method. By analyzing the experimental results of the group 5, it is concluded that adding the LSTM to deal with

TABLE 2. Experimental results on the real-set.

Method	Average rate	High	Low
HMP [60]	91.60%	93.10%	90.10%
CNN and SVM [61]	88.90%	91.80%	86.00%
CNN and RNN [62]	93.80%	96.70%	90.90%
D3D-LSTM	96.40%	99.10%	93.70%

the global long-time features can improve the performance of the model, which also proves that the idea of combining the LSTM is correct. By comparing the experimental results of groups 5 and 6, it is also possible to derive the introduction of the attention mechanism, which can improve the ability of the LSTM to deal with global long-time features. The experimental results of group 6 are better than the other 5 groups, which indicates that the D3D-LSTM model is effective and better than other traditional methods.

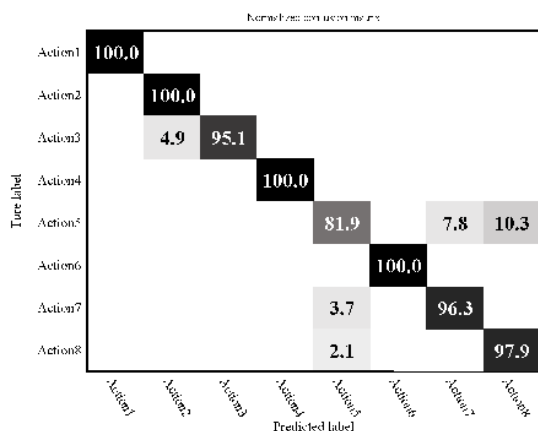


FIGURE 12. Confusion matrix of the proposed model on the Real-set.

Additionally, the confusion matrix of the proposed D3D-LSTM model on the Real-set is shown in Fig.12. By analyzing, it can be concluded that although the model occasionally misjudges the recognition of interactive actions, the overall performance is superior. Because, in the process of double target approach, especially when the target overlaps, there would be limb contact and large area occlusion, so the model can occasionally misclassify. The recognition rate for a single target action is close to 100%, that for interaction actions with a long duration is above 95%.

In summary, the proposed D3D-LSTM model is robust and feasible. However, the difference between the model and several classic algorithms is unknown. Therefore, using the Real-set to test several classical target recognition approaches, the experimental results are shown in Table 2. The results prove the effectiveness and superiority of the proposed D3D-LSTM method.

C. HUMAN ACTION RECOGNITION: SBU-KINECT

These sets of experiments intend to analyze the performance of the proposed D3D-LSTM model on recognizing interactions. The interaction is more complicated and is a key

TABLE 3. Experimental results on the SBU-Kinect dataset.

Method	Average rate
Velocity features [66]	48.40%
MIL [67]	73.80%
CHARM [68]	83.90%
SVM + RBF [69]	86.9%
CFDM + Skeleton [65]	89.4%
3D MTG [58]	90.40%
Skeleton + LSTM [64]	90.5%
LSTM [63]	91.5%
D3D-LSTM	92.40%

role in human action recognition. Each action in the dataset is long-lasting and is finished by many participants, which is challenging for the model.

The SBU-Kinect is the first publicly available RGB-D interactive action dataset, including approach, leave, kick, punch, push, hug, shake hand, and exchange item. The results of the proposed D3D-LSTM model and other state-of-the-art methods on this dataset are shown in Table 3. By comparison, the recognition rate of the model is higher than other methods. Specifically, in [58] and [63], both introduced attention mechanisms into LSTM, but the recognition rate of our model is higher due to the improved attention model. These results show that our model can extract more effective spatiotemporal features. Besides, by comparing [64] and [65], it can be obtained that our model is superior to the skeleton-based methods in terms of recognition rate. It should be noted that the methods based on different data types have their advantages. As shown in Fig.13, the confusion matrix on the SBU-Kinect is given. It can be intuitively concluded that in addition to the extremely similar behaviours, including showing, hugs, and hits (the body shape and action sequence are almost the same), the recognition rate of other interactions is about 95%.

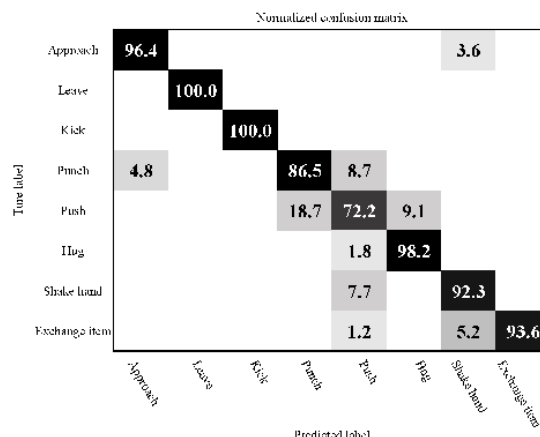


FIGURE 13. Confusion matrix of the new fusion model on the SBU-Kinect.

The improved attention mechanism introduced in the proposed D3D-LSTM model enhances its ability to extract long-term features. The unique advantage of the 3D-CNN network

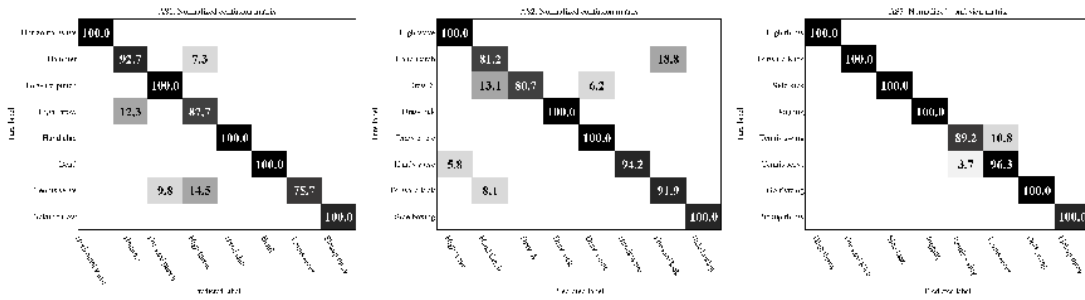


FIGURE 14. Confusion matrices of the proposed D3D-LSTM model on the MSR-action-3D dataset. AS1 and AS2: Similar actions, that is, the timing and amplitude of the actions are similar, used to test the model’s ability to distinguish similar actions. AS3: Complex actions, including actions combined with simple actions and long-lasting actions, are used to test the model’s ability to recognize complex actions.

is that it can extract local short-term features. The proposed D3D-LSTM model combines long-short-time features when processing data. This dramatically enhances the recognition rate for long-lasting behaviours, such as complex actions and interactions.

D. HUMAN ACTION RECOGNITION: MSR-ACTION-3D

These sets of experiments intend to analyze the performance of the model on distinguishing similar actions and recognizing more complex actions. Both are the most common types of actions in daily life and are key indicators for evaluating whether a model has applicability.

The MSR-action-3D is a large dataset containing 20 individual actions. Following the complexity and similarity, these actions are divided into three groups, each group contains 8 actions, where high-injection, croquet and throwing, tennis serve, and front kick are simultaneously present in different groups. As shown in Table 4, the first group (AS1) and the second group (AS2) include some similar actions, and the third group (AS3) includes some more complex actions. The biggest advantage of using the dataset is that it can comprehensively evaluate the performance of the model, that is, the accuracy of distinguishing similar actions and the recognition rate of complex actions.

TABLE 4. MSR-action-3D dataset.

AS1	AS2	AS3
Horizontal wave	High wave	High throw
Hammer	Hand Catch	Forward kick
Forward punch	Draw X	Side kick
Forward punch	Draw circle	Jogging
Hand clap	Draw tick	Tennis swing
Bend	Hands wave	Tennis serve
Tennis serve	Side boxing	Golf swing
Pickup throw	Forward kick	Pickup throw

The experimental results of the proposed D3D-LSTM model and other state-of-the-art methods are shown in Table 5. By comparison, it can be concluded that the recognition rate of complex actions in the model is better than other methods, and the recognition rate of most similar

TABLE 5. Experimental results on the MSR-action-3D dataset.

Method	Average rate	AS1	AS2	AS3
HOD [57]	91.30%	92.40%	90.20%	91.40%
STOP [70]	87.50%	91.70%	72.20%	98.60%
3D MTG [58]	94.40%	92.40%	93.80%	97.10%
Deep model [14]	93.30%	90.80%	93.40%	95.70%
D3D-LSTM	95.40%	94.50%	92.90%	98.90%

actions is also higher than other methods. This is because, in the new model, the real-time feature fusion method can extract more effective space features without losing important information, thus improving the ability to distinguish similar actions. The combination of 3D-CNN and LSTM, and the introduction of an improved attention mechanism, which allows the model to better process temporal information, thereby improving the recognition performance of complex actions.

Figure 14 shows the confusion matrices for the results of the proposed model. As shown, the recognition rate of most of the actions is as high as 90%, especially the recognition rate of complex actions in the AS3 is close to 100%. In other state-of-the-art methods, most models cannot have a high recognition rate for “ Pickup throw ” in the AS3. This is mainly because that “ Pickup throw ” is a continuous long-term combination action, that is, the ball is thrown after the ball is picked up first. However, other methods have insufficient ability to process the global long-time feature.

In the AS1 and AS2, in addition to these very similar actions, the recognition rate of other actions is also close to 100%.The very similar actions recognition rate is also stable at 90% which is better than other state-of-the-art methods. For example, in previous studies, the recognition rate of “Pickup throw” and “Bend” in the AS1 are not good. Especially, it is noticed that the recognition rate of our model in the AS2 is not the best, the main reasons are as followed. In [58], the 3DMTG model was proposed for skeleton joints feature extraction based on a new histogram projection method and a novel feature descriptor. The main contribution is to improve the ability to distinguish similar actions by recording 3D moving trend feature in body joints. However, since the

model did not fuse RGB and depth features, the recognition rate for complex, angle-changing or partially occluded actions is not better than the D3D-LSTM model, and cannot be widely applied in the real world. In [14], spatiotemporal multi-fused features-based an online HAR method was proposed for depth and skeleton data action recognition. The method can track body parts such as arms and legs in case of multiple actions, which aims to describe in detail the differences between similar activities. While the method is able to detect minor differences in action, the modeling ability of complex spatial features is not better than our model due to the lack of color information.

TABLE 6. Quantitative evaluation of HMM-based methods on the MSR-action-3D dataset.

Method	Average rate	AS1	AS2	AS3
HMM [71]	83.92%	/	/	/
HMM & GMM [14]	93.30%	90.80%	93.40%	95.70%
HMM & SOM [6]	82.10%	81.70%	82.50%	81.20%
HMM & SVM [72]	89.10%	/	/	/
D3D-LSTM	95.40%	94.50%	92.90%	98.90%

We compare our model with state-of-the-art HMM-based models on the MSR-action-3D dataset and summarize the results in Table 6. The hidden Markov model (HMM) achieves better ability to model temporal feature and have been extensively used in sequence recognition. We consider the following state-of-the-art methods: Jalal *et al.* developed the multi-fused features coding method for training the HMM model [14], and trained clustered features based on transition and emission probabilities values [6], and extended HMM by using robust depth silhouettes context features [71]; Wu *et al.* combined SVM and HMM for continuous action feature modeling [72].

In summary, the proposed D3D-LSTM model achieves great advantages in recognizing complex actions, and the performance of distinguishing very similar actions is better than other methods.

VI. CONCLUSION AND FUTURE WORK

We propose the D3D-LSTM model for recognizing human action based on RGB-D. The proposed D3D-LSTM model is based on 3D-CNN and LSTM, which also introduces the idea of dense connection, the improved attention mechanism, and the real-time feature fusion method. The model has a strong global long-term feature processing performance and can extract better spatiotemporal features which increase the recognition rate of complex actions as well as distinguish similar actions. We collect a dataset called Real-set with changing scenes, which currently is a more realistic RGB-D action dataset.

A series of experiments are conducted to compare the proposed D3D-LSTM model with other traditional methods, to prove the correctness of the study ideas. Because the model improves the extracting ability of global features. The performance of the model in the Real-set, SBU-Kinect, and MSR-action-3D data sets is superior to other state-of-the-art

methods. Especially, the recognition rate of complex actions in MSR-action-3D is about 5% higher than other methods, and the average accuracy rate is improved by 2% when distinguishing similar actions. When recognizing interactions in SBU-Kinect, the recognition rate is increased by about 3%. These databases are challenging because they contain similar actions, complex actions, and multiple changes. However, some limitations should be noted. First, the D3D-LSTM model has not achieved the best recognition rate for intra-class similarity actions, such as *Tennis serve*, *Draw X*, *Hand catch*. Second, our model can extract the keyframes of each sample, but it cannot automatically extract saliency information in the keyframes that would be important for the recognition rate.

In future work, we will improve the effectiveness of distinguishing similar actions by adding the skeleton data and research a more discriminative temporal attention model. Besides, we also focus on continuous action recognition to make the proposed method more practical in real applications.

REFERENCES

- [1] H.-Y. Cheng and J.-N. Hwang, "Integrated video object tracking with applications in trajectory-based event detection," *J. Vis. Commun. Image Represent.*, vol. 22, no. 7, pp. 673–685, Oct. 2011.
- [2] A. Y. Yang, S. Iyengar, P. Kuryloski, and R. Jafari, "Distributed segmentation and classification of human actions using a wearable motion sensor network," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–8.
- [3] A. Jalal and S. Kamal, "Real-time life logging via a depth silhouette-based human activity recognition system for smart home services," in *Proc. 11th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2014, pp. 74–80.
- [4] Y. Song, J. Tang, F. Liu, and S. Yan, "Body surface context: A new robust feature for action recognition from depth videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 952–964, Jun. 2014.
- [5] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.
- [6] A. Jalal, S. Kamal, and D. Kim, "Shape and motion features approach for activity tracking and recognition from kinect video camera," in *Proc. IEEE 29th Int. Conf. Adv. Inf. Netw. Appl. Workshops*, Mar. 2015, pp. 445–450.
- [7] A. Jalal, S. Kamal, and D. Kim, "A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments," *Sensors*, vol. 14, no. 7, pp. 11735–11759, Jul. 2014.
- [8] K. Buys, C. Cagniat, A. Baksheev, T. De Laet, J. De Schutter, and C. Pantofaru, "An adaptable system for RGB-D based human body detection and pose estimation," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 39–52, Jan. 2014.
- [9] A. Jalal, J. T. Kim, and T.-S. Kim, "Human activity recognition using the labeled depth body parts information of depth silhouettes," in *Proc. 6th Int. Symp. Sustain. Healthy Buildings*, Seoul, South Korea, vol. 27, 2012, pp. 1–8.
- [10] J. Puwein, L. Ballan, R. Ziegler, and M. Pollefeys, "Joint camera pose estimation and 3D human pose estimation in a multi-camera setup," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 473–487.
- [11] K. Nour el houda Slimani, Y. Benezeth, and F. Souami, "Human interaction recognition based on the co-occurrence of visual words," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 455–460.
- [12] X. Zhao, X. Li, C. Pang, and S. Wang, "Human action recognition based on semi-supervised discriminant analysis with global constraint," *Neuro-computing*, vol. 105, pp. 45–50, Apr. 2013.
- [13] A. Jalal, Y. Kim, and D. Kim, "Ridge body parts features for human pose estimation and recognition from RGB-D video data," in *Proc. 5th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2014, pp. 1–6.

- [14] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern Recognit.*, vol. 61, pp. 295–308, Jan. 2017.
- [15] A. Jalal, Y. Kim, S. Kamal, A. Farooq, and D. Kim, "Human daily activity recognition with joints plus body features representation using kinect sensor," in *Proc. Int. Conf. Inform., Electron. Vis. (ICIEV)*, Jun. 2015, pp. 1–6.
- [16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: <http://arxiv.org/abs/1502.03167>
- [19] 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.
- [20] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.
- [21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
- [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.* Cham, Switzerland: Springer, 2016, pp. 20–36.
- [23] Z. Li, K. Gavriluyk, E. Gavves, M. Jain, and C. G. M. Snoek, "VideoLSTM convolves, attends and flows for action recognition," *Comput. Vis. Image Understand.*, vol. 166, pp. 41–50, Jan. 2018.
- [24] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
- [25] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1583–1597, Aug. 2016.
- [26] 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.
- [27] Z. Shi and T.-K. Kim, "Learning and refining of privileged information-based RNNs for action recognition from depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3461–3470.
- [28] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. Ogunbona, "ConvNets-based action recognition from depth maps through virtual cameras and pseudocoloring," in *Proc. 23rd ACM Int. Conf. Multimedia (MM)*, 2015, pp. 1119–1122.
- [29] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 667–681, Mar. 2018.
- [30] H. Rahmani and A. Mian, "3D action recognition from novel viewpoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1506–1515.
- [31] C. J. Dhamsania and T. V. Ratanpara, "A survey on human action recognition from videos," in *Proc. Online Int. Conf. Green Eng. Technol. (IC-GET)*, Nov. 2016, pp. 1–5.
- [32] S. J. Berlin and M. John, "Human interaction recognition through deep learning network," in *Proc. IEEE Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2016, pp. 1–4.
- [33] C. Chattopadhyay and S. Das, "Supervised framework for automatic recognition and retrieval of interaction: A framework for classification and retrieving videos with similar human interactions," *IET Comput. Vis.*, vol. 10, no. 3, pp. 220–227, Apr. 2016.
- [34] A. Farooq, A. Jalal, and S. Kamal, "Dense RGB-D map-based human tracking and activity recognition using skin joints features and self-organizing map," *KSI Trans. Internet Inf. Syst.*, vol. 9, no. 5, pp. 1856–1869, 2015.
- [35] S. Kamal, A. Jalal, and D. Kim, "Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM," *J. Elect. Eng. Technol.*, vol. 11, no. 6, pp. 1857–1862, Nov. 2016.
- [36] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
- [37] 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.
- [38] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Multi-fiber networks for video recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 352–367.
- [39] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank, "Asymmetric 3D convolutional neural networks for action recognition," *Pattern Recognit.*, vol. 85, pp. 1–12, Jan. 2019.
- [40] N. Hussein, E. Gavves, and A. W. M. Smeulders, "Timeception for complex action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 254–263.
- [41] C.-Y. Wang, C.-C. Chiang, J.-J. Ding, and J.-C. Wang, "Dynamic tracking attention model for action recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1617–1621.
- [42] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.
- [43] X. Yang and Y. Tian, "Effective 3D action recognition using Eigen-Joints," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 2–11, Jan. 2014.
- [44] M. Li and H. Leung, "Graph-based approach for 3D human skeletal action recognition," *Pattern Recognit. Lett.*, vol. 87, pp. 195–202, Feb. 2017.
- [45] J. Luo, W. Wang, and H. Qi, "Spatio-temporal feature extraction and representation for RGB-D human action recognition," *Pattern Recognit. Lett.*, vol. 50, pp. 139–148, Dec. 2014.
- [46] X. Jiang, F. Zhong, Q. Peng, and X. Qin, "Action recognition based on global optimal similarity measuring," *Multimedia Tools Appl.*, vol. 75, no. 18, pp. 11019–11036, 2016.
- [47] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 525–538.
- [48] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 804–811.
- [49] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
- [50] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.
- [51] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: <http://arxiv.org/abs/1409.0473>
- [52] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.
- [53] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," 2014, *arXiv:1412.7755*. [Online]. Available: <http://arxiv.org/abs/1412.7755>
- [54] S. Sharma, R. Kiro, and R. Salakhudinov, "Action recognition using visual attention," 2015, *arXiv:1511.04119*. [Online]. Available: <http://arxiv.org/abs/1511.04119>
- [55] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
- [56] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," 2014, *arXiv:1412.4729*. [Online]. Available: <http://arxiv.org/abs/1412.4729>
- [57] Z. Huang, C. Wan, T. Probst, and L. V. Gool, "Deep learning on lie groups for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6099–6108.
- [58] B. Liu, H. Yu, X. Zhou, D. Tang, and H. Liu, "Combining 3D joints moving trend and geometry property for human action recognition," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2016, pp. 332–337.

[59] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2752–2759.

[60] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for RGB-D based object recognition," in *Experimental Robotics*. Heidelberg, Germany: Springer, 2013, pp. 387–402.

[61] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 345–360.

[62] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3D object classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 656–664.

[63] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017.

[64] F. Baradel, C. Wolf, and J. Mille, "Pose-conditioned spatio-temporal attention for human action recognition," 2017, *arXiv:1703.10106*. [Online]. Available: <https://arxiv.org/abs/1703.10106>

[65] Y. Ji, H. Cheng, Y. Zheng, and H. Li, "Learning contrastive feature distribution model for interaction recognition," *J. Vis. Commun. Image Represent.*, vol. 33, pp. 340–349, Nov. 2015.

[66] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013.

[67] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 28–35.

[68] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "On the improvement of human action recognition from depth map sequences using space–time occupancy patterns," *Pattern Recognit. Lett.*, vol. 36, pp. 221–227, Jan. 2014.

[69] 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.

[70] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[71] A. Jalal, S. Kamal, and D. Kim, "Individual detection-tracking-recognition using depth activity images," in *Proc. 12th Int. Conf. Ubiquitous Robots Ambient Intell. (URAI)*, Oct. 2015, pp. 450–455.

[72] H. Wu, W. Pan, X. Xiong, and S. Xu, "Human activity recognition based on the combined SVM&HMM," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Jul. 2014, pp. 219–224.



WEI YANG was born in Heilongjiang, China, in 1994. He received the B.S. degree from Shenyang Ligong University, China, in 2016. He is currently pursuing the M.S. degree with Shenyang Ligong University. His research interests include machine learning and digital image processing.



YUEQIU JIANG received the Ph.D. degree in computer application technology from Northeastern University, in 2004. Since 2010, she has been a Full Professor with Shenyang Ligong University. She is currently the Leader of the subject direction for signal and information process. Her research interests include network management and image processing.



WEIHONG CHIN received the bachelor's degree (Hons.) in electronics engineering, majoring in robotics and automation from Multimedia University, Malacca, Malaysia, in 2011, the master's degree in computer science from the University of Malaya, Malaysia, in 2015, and the Ph.D. degree from Tokyo Metropolitan University, Japan. He is currently an Assistant Professor with Tokyo Metropolitan University. His current research interests include biologically inspired incremental learning, robot navigation, and human–robot interaction.



NAOYUKI KUBOTA received the B.Sc. degree from Osaka Kyoiku University, Kashiwara, Japan, in 1992, the M.Eng. degree from Hokkaido University, Hokkaido, Japan, in 1994, and the D.E. degree from Nagoya University, Nagoya, Japan, in 1997.

He joined the Osaka Institute of Technology, Osaka, Japan, in 1997. He joined the Department of Human and Artificial Intelligence Systems, University of Fukui, Fukui, Japan, as an Associate Professor, in 2000. He joined the Department of Mechanical Engineering, Tokyo Metropolitan University, Tokyo, in 2004. He was an Associate Professor (2005–2012), and has been a Professor with the Department of System Design, Tokyo Metropolitan University, since 2012.



ZHAOJIE JU (Senior Member, IEEE) received the B.S. degree in automatic control and the M.S. degree in intelligent robotics from the Huazhong University of Science and Technology, China, and the Ph.D. degree in intelligent robotics from the University of Portsmouth, U.K.

He held a research appointment at the University College London, London, U.K., before he started his independent academic position at the University of Portsmouth, U.K., in 2012. His research interests include machine intelligence, pattern recognition, and their applications on human motion analysis, multifingered robotic hand control, human–robot interaction and collaboration, and robot skill learning. He has authored or coauthored over 180 publications in journals, book chapters, and conference proceedings and received four best paper awards and one Best AE Award in ICRA2018.



JIAHUI YU received the B.S. and M.S. degrees, majoring in intelligent systems, from Shenyang Ligong University, China, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree with the University of Portsmouth, U.K. His current research interests include machine intelligence, pattern recognition, and human-robot/computer interaction and collaboration.



HONGWEI GAO received the Ph.D. degree in the field of pattern recognition and intelligent system from the Shenyang Institute of Automation (SIA), Chinese Academy of Sciences (CAS), in 2007. Since September 2015, he has been a Professor with the School of Automation and Electrical Engineering, Shenyang Ligong University. He is currently the Leader of academic direction for optical and electrical measuring technology and system. His research interests include digital

image processing and analysis, stereo vision, and intelligent computation. He has published more than sixty technical articles in these areas as the first author or coauthor.

...