A Two-Stream Recurrent Network for Skeleton-based Human Interaction Recognition

Qianhui Men*, Edmond S. L. Ho†, Hubert P. H. Shum‡, and Howard Leung*

*Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China
†Department of Computer and Information Sciences, Northumbria University, Newcastle, UK
‡Department of Computer Science, Durham University, Durham, UK

Abstract—**This paper addresses the problem of recognizing human-human interaction from skeletal sequences. Existing methods are mainly designed to classify single human action. Many of them simply stack the movement features of two characters to deal with human interaction, while neglecting the abundant relationships between characters. In this paper, we propose a novel two-stream recurrent neural network by adopting the geometric features from both single actions and interactions to describe the spatial correlations with different discriminative abilities. The first stream is constructed under pairwise joint distance (PJD) in a fully-connected mesh to categorize the interactions with explicit distance patterns. To better distinguish similar interactions, in the second stream, we combine PJD with the spatial features from individual joint positions using graph convolutions to detect the implicit correlations among joints, where the joint connections in the graph are adaptive for flexible correlations. After spatial modeling, each stream is fed to a bi-directional LSTM to encode two-way temporal properties. To take advantage of the diverse discriminative power of the two streams, we come up with a late fusion algorithm to combine their output predictions concerning information entropy. Experimental results show that the proposed framework achieves state-of-the-art performance on 3D and comparable performance on 2D interaction datasets. Moreover, the late fusion results demonstrate the effectiveness of improving the recognition accuracy compared with single streams.**

I. Introduction

Modeling human-human interactions in competitive sports (such as *kick-boxing* [1]) and close interactions with tangling limbs (such as *judo* [2]) have benefited the computer animation community. Enhancing the motion features such as interaction mesh [2] demonstrate better results in motion retrieval [3], [4] and classification tasks [5]. More recently, the emergence of Recurrent Neural Network (RNN) enables automatically modeling the temporal dependency of motion sequences. As a reliable input, hand-crafted features may help RNN-based networks better learn spatial reasoning. For example, different types of geometric features from a single human are proved to be more effective than raw joints under a stack Long Short-Term Memory (LSTM) network [6]. The hand-crafted features are also useful in the skeleton-based interaction recognition, as their rich geometric information may indicate spatial dependency within the character(s).

However, previous works in interaction recognition [7], [8] mainly focus on extracting features from individual characters that ignore the contextual information of the interaction between characters. Another problem is the lack of exploring the correlations among joints, such as modeling body part relations [5], or plainly stacking the joint features [9], which causes the similar interactions less likely to be distinguished by such models. To this end, we propose a two-stream framework by exploring interaction representations under distance-based and joint-based features, respectively. Each of the streams demonstrates a strong discriminative power for certain types of human interactions: In the first stream, we adopt pairwise joint distance (PJD) as geometric features within two characters to effectively classify the interactions with explicit patterns of relative distances, such as *kicking* and *punching*; In the second stream, we design graph convolutions on joint positions to learn their implicit correlations with the spatial proximity represented by the pairwise distance features, which is effective in classifying interactions with similar PJD features but different joint movements, such as *pushing* and *punching*; Finally, we fuse their recognition outputs to take advantage of the different discriminative abilities of two streams.

To leverage both individual and interaction features, we construct the first stream as a bi-directional LSTM (BiLSTM) network with the intra-subject and inter-subject pairwise joint distances as the geometric features to discriminate interactions. The effectiveness of PJD was demonstrated in modeling motion sequences performed by a single subject [10]. Such an approach can also be used to explicitly characterize an interaction. For example, in *shaking hands*, we always observe a smaller PJD between two hands of the two characters. We thus model PJD between every joint pair of two subjects, which ends up with a fully-connected mesh to spatially describe an interaction. The bi-directional LSTM further equips the network with long term dependencies to minimize the information loss problem in modeling long sequences. By encoding the two-way temporal information, the recognition performance can be further improved.

To explore the correlations among joints, we then propose the second stream of a fully-connected graph convolution BiLSTM network with adaptive joint connectivity. The spatial proximity of interaction is represented using fully-connected PJD defined on this adaptive graph. Graph convolutional network (GCN) [11] has become an effective tool for analyzing joint correlations within a single action sequence under graphical structure [12], [13]. Here, we leverage the graph convolution to model the interdependency among joints in two-character interactions, and the spatial proximity of interaction is represented by the fully-connected PJD from the first stream since it can be naturally extended to describe the strengths of edges in the graph. Previous graph convolutions on joints [12], [13] limit the motion representations under fixed graph connectivity (i.e., the kinematic tree), which spatially restrict the graph to a predefined topological structure. In our graph topology, we make the graph connectivity to be adaptive to allow flexible connections between joints of two characters, which aims to highlight the important joint pairs from the abundant information. Temporally, we also embed the designed fully-connected graph convolution into a bi-directional LSTM network to encode two-way interaction dynamics.

We further propose a late fusion algorithm to improve the recognition performance of the proposed hybrid networks. Previous work conducted on hierarchical networks, such as spatial-temporal Convolutional Neural Networks (CNNs) [14], two-stream RNNs [9], and multi-clip skeleton images [15], tend to average the prediction scores from individual models. However, the classification power of models may vary due to different spatial-temporal representations. To achieve an optimal score, we link the classification posterior probabilities with its information entropy to produce a sparse mixture distribution as the optimal estimator.

In summary, the main contributions are concluded as follows:

- We propose a pairwise joint distance BiLSTM network (PJD-BiLSTM) that models the explicit interaction patterns from the discriminative geometric features within two characters.
- We propose a fully-connected graph convolution BiLSTM network (FCGC-BiLSTM) that quantifies the spatial proximity of interaction from both joint positions and PJD features with adaptive graph connections to extract the implicit correlations among joints.
- A late fusion algorithm is defined to boost the recognition accuracy from probability outputs of the proposed network streams.

Experimental results show that our method outperforms the state of the arts on the 3D skeleton dataset. We also demonstrate our method can be easily extended to 2D key joint recognition and achieves comparable performance with RGB-based methods.

II. Related Work

In this section, we first review the conventional approaches to human interaction modeling. Then we recall the existing neural network models which are closely related to our proposed framework.

A. Skeleton-based Interaction Modeling

Modeling and synthesizing two-character interactions based on skeleton structure have been explored in the area of computer vision and computer graphics. Early work by Shum et al. [1] computed the basic patterns of two-character interactions in *kick-boxing* as interaction patches based on game-tree expansion [16]. For motion comparison, Tang et al. [10] proposed to use all combinations of relative joint distances as features with feature selection algorithms for content-based motion retrieval. Yun et al. [17] evaluated a wide range of motion features, such as joint positions, velocity, relational features for classifying human-human interactions captured from the Kinect sensor.

The interaction mesh [2] is effective in modeling the spatial relationship between different body parts for motion retargeting applications. The concept of interaction mesh was further extended in proximity graph [4] to compute the similarity between different human-human interactions. One problem of using interaction mesh for interaction comparison is the discrete nature of the mesh construction. As a result, small changes in the 3D position will result in a different mesh topology. To alleviate this problem, more recent approaches [5], [18] proposed sampling a very dense set of points from the characters for interaction classification and more generic comparison tasks.

B. Neural Networks in Action Recognition

RNN and its variants show potentials in modeling temporal dynamics, which rises great attention for human activity analysis in recent research. Zhu et al. [19] learned the intrinsic co-occurrence of joints with LSTM neurons. Zhang et al. [20] designed an adaptive RNN network which can regulate the viewpoints to maximize recognition accuracy. However, RNN-based networks may fail to detect the spatial patterns by simply concatenating the skeleton joints into a chain sequence [20], [21]. Therefore, based on RNN frameworks, some researches [7], [22] proposed to describe the spatial path along a graph-based kinematic tree. Others [9], [23] proposed hierarchical networks for several branches of the human skeleton (normally four limbs and a trunk) and gathered the output representations under a fine-to-coarse fashion.

Another solution to identify the latent correlations among joints is to model actions using CNN or GCN. In [15], joint positions were transferred to skeleton images and then fed to a multi-stage CNN to learn the spatial-temporal properties. Tas and Koniusz [24] created feature maps of actions using kernel linearization. Later, GCN demonstrated a great advantage in human action recognition where the skeleton joints as nodes interact with each other in graph format. Yan et al. [12] performed graph convolution by constructing edges from natural joint connectivities and consecutive frames along with spatial and temporal directions, respectively. There are also some cutting-edge researches associating GCN with RNN to jointly learn the spatial-temporal correlations. For example, Si et al. [25] inserted graph convolutions in LSTM cells together with an attention mechanism to enhance the informative joints or body parts. However, these architectures exploit GCN mainly based on raw skeleton connections. In this paper, we incorporate GCN with pairwise geometric features

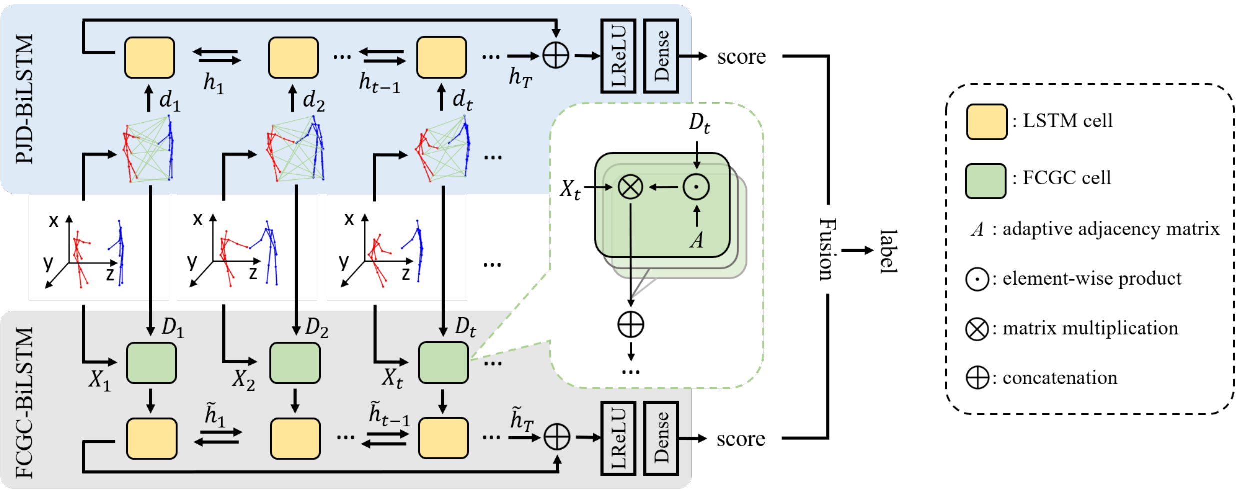


Fig. 1. The unrolled version of our proposed two-stream framework. The FCGC structure is detailed illustrated in the callout box enclosed by the green dotted line.

under adaptive graph connections to explore more robust correlations among joints.

III. Methodology

The overview of our proposed framework is illustrated in Fig. 1. It consists of two streams, namely Pairwise Joint Distance BiLSTM (PJD-BiLSTM, Section III-B) and Fully-Connected Graph Convolution BiLSTM (FCGC-BiLSTM, Section III-C). The input skeletal features include the pairwise joint distances and joint positions for different streams which are effective in modeling the spatial relations of the two-character and individual postures respectively. For example, pairwise joint distance is better at discriminating explicit interaction patterns such as leg movements of *kicking* and arm movements of *punching*, while joint positions can discover implicit correlations among joints that help discriminate similar interactions like *punching* and *pushing*. We then present our late fusion method by combining the two separate predictions for the best outcome.

A. Problem Formulation

We first parameterize human interaction with the joint position and PJD. The PJD feature [10] shows an advantage in modeling spatial proximity between joints which explicitly characterizes the interaction. Given a skeletal interaction sequence, the input joint positions can be denoted as $\mathbf{X} = [X_1, X_2, ..., X_T]$, where T is the total number of motion frames. In each frame, X_t contains the joint positions of the two subjects with C channels. Note that $C = 2$ for 2D pose and $C = 3$ for 3D skeleton.

The PJD between the two subjects is computed to represent the interaction, and it can also be used as edge information when a graph-based model is applied [26]. Following [17] which extracted the Euclidean distance between joints of the two subjects, the PJD between joint i in X^p and joint j in X^q is defined as:

$$d(X_i^p, X_j^q) = \|X_i^p - X_j^q\|, \tag{1}$$

where i and j are in set J that contains all the joints of a subject. Note that the distance is computed from the two subjects if $p \neq q$, and from a single subject if $p = q$. PJD will play different roles in each separate stream which demonstrates different advantages in the interaction recognition task.

B. Pairwise Joint Distance BiLSTM

As the first stream, we propose a pairwise joint distance BiLSTM framework to learn the spatial-temporal dependencies characterized by PJD. The stream aims at modeling the spatial correlations of interaction from distance space with all the joint pairs of two characters.

We first generate a mesh for human-human interaction with each of the edges connecting any of the two joints, such that it captures not only features within a single character but also between two characters. The left part of Fig. 2 illustrates all the connected joint pairs in an interaction between a simplified humanoid skeletal structure with five joints. The changes in edge lengths (i.e., deformation of the whole mesh) over time represent the dynamics of spatial proximity within the interaction.

As shown in the upper part of Fig. 1, we feed the fully-connected PJD within the mesh into a BiLSTM layer in forward and backward order to learn the temporal properties of the spatial proximity. The bi-directional scheme builds a context-aware model by reducing the long-term information loss, which enhances the classification performance for sequences (see the accuracy improvement in Table III). The final concatenation of the bi-directional outputs goes through a dense layer with softmax activation for probability forecasting towards each interaction label. With the help of the interaction features, using one layer BiLSTM could already achieve

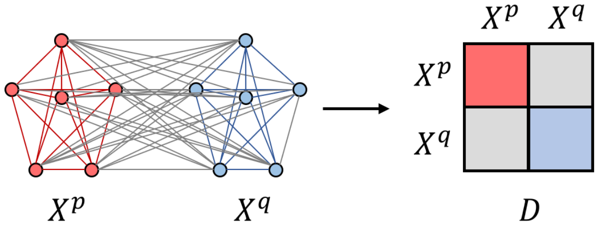


Fig. 2. The transformation from fully-connected mesh to weight matrix in FCGC.

results comparable to the state-of-the-art approaches as shown in Section IV especially when the evaluated dataset has a small scale.

C. Fully-Connected Graph Convolution BiLSTM

In the second stream, we propose a fully-connected graph convolution BiLSTM framework to incorporate the distance metric PJD with the corresponding joint positions in a complementary way to learn the latent spatial properties among joints. From the fact that PJD may overlook the joint distributions which contain representative spatial patterns, we use extra positional features to complementary learn the correlations among joints. Since human interaction can be naturally represented as graph-structured data with the nodes formed by joints, we further adopt graph convolutions [11] to spatially model the correlations among joint positions based on the spatial proximity learned from PJD.

We quantify the spatial proximity in our graph structure with PJD under adaptive joint connectivity. In graph-based models, a node can gather the information from the other nodes according to the spatial proximity of the graph, which is represented by an adjacency matrix with each entry denoting whether the corresponding two nodes are connected in the graph. Previous models [12], [13] using graph representations are always modeling single actions, and their adjacency matrices only involve the joint connections based on the kinematic tree. In a recent work of action prediction, [27] made their adjacency matrix to be adjustable with network training, which shows a better performance than using the predefined skeletal structure based on bone connections. Here, we partially adopt their idea by using an adaptive matrix denoting the importance of joint connectivity but combining with the fully-connected PJD to support the spatial proximity. This is because PJD holds an advantage to show the joint relationship that two close joints are highly correlated and vice versa. Here, PJD, together with the adaptive connectivity functions as the adjacency matrix allowing flexibility to describe the spatial proximity of the graph.

The human-human interaction is presented as a fully-connected graph g with $2|J|$ nodes corresponding to all joints of the two characters. We assume that two joints are highly interacted if they are in close proximity. To obtain the spatial proximity, we first transfer the PJD into matrix format D as shown in the right part of Fig. 2. Note that the weighted distance matrix D is symmetric. More specifically, at a certain timestamp t, the connection strength $D_t(X_i^p, X_j^q)$ between two nodes X_i^p and X_j^q will be calculated by taking the reciprocal of exponential PJD as

$$D_t(X_i^p, X_j^q) = \frac{1}{\exp(d_t(X_i^p, X_j^q))} \tag{2}$$

to normalize its scope within $(0, 1]$. This step ensures that a joint pair with small PJD will have a large weight in D.

To determine the physical intensities of the pairwise connections, we propose an adaptive connection map $A \in \mathbb{R}^{2|J|\times 2|J|}$ with trainable parameters. This avoids manually defining the fixed connections within the kinematic tree. As a data-driven approach, an adaptive A will also capture the implicit connections of the joint pairs between two characters. For example, the connection of two hands is important in *clapping*, but the hands are not directly connected in the kinematic chain.

Given all joint positions $\mathbf{X}_t = [X^p; X^q]|_t$ as input, the fully-connected graph convolution (FCGC) operation (the green module in Fig. 1) under a set of convolutional coefficients W can be written as:

$$W \star_g \mathbf{X}_t = \sigma((\overset{C}{\underset{c=1}{\oplus}}(A \odot D_t)\mathbf{X}_t^C)W), \tag{3}$$

where $\sigma(\cdot)$ denotes the activation function, and each element of the dot product $A \odot D_t$ captures the location dependencies of the corresponding joint pairs. Note that $A \odot D_t$ is time specific and its temporal variations indicate the changing of spatial proximity of the certain interaction over time. Rather than directly encoding the absolute coordinate $\mathbf{X}_t$, the multiplication step with the dot product equips the joint coordinate with global relationship among the other joints.

We further embed FCGC with LSTM cell at each frame t to encode the temporal dynamics:

$$\tilde{h}_t = LSTM([\star_g \mathbf{X}_t, \tilde{h}_{t-1}]; [W_x, U_{\tilde{h}}]), \tag{4}$$

where $\star_g$ is the graph convolution operator defined in Eq. (3), the convolutional coefficients W becomes W_x which is the LSTM weights of input $\mathbf{X}_t$ for each gate, and $U_{\tilde{h}}$ is the weights for the hidden state $\tilde{h}_{t-1}$. Here, the activation function in Eq. (3) becomes $tanh(\cdot)$ as defined in a standard LSTM unit. We perform FCGC with the input $\mathbf{X}$ but not for hidden unit $\tilde{h}$, since $\tilde{h}$ does not contain explicit interactive meanings. The FCGC also goes through a backward LSTM layer in reversed time order as PJD-BiLSTM to learn the two-way temporal representations.

D. Late Fusion

To take advantage of both streams, we propose to merge the prediction score at the probability level with a scalable late fusion approach to highlight the more discriminative stream prediction based on each interaction instance. This is because the proposed two-stream networks present particular discriminative power among various interaction classes. For example, PJD-BiLSTM is good at modeling the relative position changes between two characters, while FCGC-BiLSTM is more effective when two interactions have similar PJD, such as *push* and *punch* (see Fig. 3 and Fig. 4) since the extra joint correlations are captured to differentiate such similar interactions. Compared to combining different models at an early stage, late fusion at score level considers the diversity between classifiers. Empirically, the fusion algorithm effectively combines information from both streams and improves the overall classification accuracy (see Table I and Table II).

Inspired by the Principle of Maximum Entropy [28], we re-weigh the output probability of each stream according to their entropy magnitude. Given an interaction entry $\mathbf{X}$, the n-th classifier will generate a score distribution $P_n(y_k|\mathbf{X}, \Theta_n)|_{k=1}^{K} \in [0, 1]$ among all K interaction classes, and Θ_n represents the parameter set of classifier n. We calculate the final fusion score by the following weighted average equation:

$$P(y_k|\mathbf{X}, \Theta_1, \dots, \Theta_N) = \sum_{n=1}^{N} \alpha_n P_n(y_k|\mathbf{X}, \Theta_n). \qquad (5)$$

Here, α_n gives the degree of confidence towards the n-th classifier, which is induced from:

$$\alpha_n = 1 - \frac{\sum_{k=1}^{K} P_n(y_k|\mathbf{X}, \Theta_n) \log(P_n(y_k|\mathbf{X}, \Theta_n))}{\sum_{m=1}^{N} \sum_{k=1}^{K} P_m(y_k|\mathbf{X}, \Theta_m) \log(P_m(y_k|\mathbf{X}, \Theta_m))}. \qquad (6)$$

The numerator refers to the negative information entropy of label distribution for the n-th classifier. Our goal is to highlight the distributions with lower entropy that indicates higher confidence of the predicted class, and to hold back the less discriminative predictions in the meanwhile. Here $N = 2$ corresponds to the two streams, and the proposed fusion algorithm can also be generalized to multiple classifiers to boost the recognition performance.

IV. Experiments

To show the effectiveness and general applicability of our two-stream recognition approach, we examine both 3D skeleton with depth and 2D skeleton with key joint positions in RGB videos. The benchmark datasets used in this study are SBU Kinect Interaction [17] for 3D, and UT-Interaction [29] for 2D skeleton-based recognition. The experiments are carried out on the proposed PJD-BiLSTM, FCGC-BiLSTM, and two-stream fusion (denoted as PJD+FCGC). We also compare with the state of the arts and perform ablation study to justify each component in the proposed network.

A. Datasets and Implementation Details

For splitting the training and testing samples, we follow the experiment protocol of [17] with 5-fold cross-validation on the SBU dataset, and 10 folds [29] on the UT-Interaction dataset. The averaged result for all folds is presented as the final accuracy for each dataset. The numbers of LSTM neurons are set to 1024 and 512 for PJD-BiLSTM and FCGC-BiLSTM, respectively. The output of each stream before the softmax layer is activated by LeakyReLU with a negative slope of 0.2. We also accommodate common techniques like gradient clipping and early stopping during network training.

TABLE I Recognition performance on the SBU dataset.

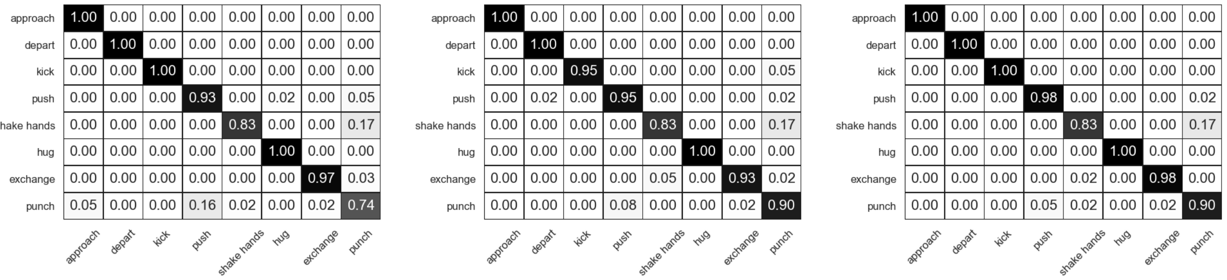
Method	Acc.(%)
Raw Skeleton [17]	49.7
Joint feature [17]	80.3
Co-occurrence LSTM [19]	90.4
ST-LSTM+Trust Gate [22]	93.3
Clips+CNN+MTLN [15]	93.5
SI and JD features [5]	93.9
GCA-LSTM [7]	94.1
CNN+Kernel Feature maps [24]	94.3
Two-stream RNN [9]	94.8
LSTM+FA+VF [8]	95.0
PJD-BiLSTM	94.0
FCGC-BiLSTM	95.1
PJD+FCGC	**96.8**

1) SBU Interaction Dataset: The SBU dataset [17] includes 282 video sequences with 8 interaction categories (i.e., *approach*, *depart*, *kick*, *push*, *punch*, *hug*, *shake hands* and *exchange*) performed by 7 participants. It also provides the identification of *active* agent and *inactive* agent for exploring and simulating. To better fit FCGC-BiLSTM, we expand the dataset by flipping the positions of active and inactive character. Note that for both streams, the pairwise distance will not change by this operation. Since the dataset shows less samples in some classes like *shake hands*, we also adopt weight balancing techniques [30] to balance all categories during training.

2) UT-Interaction Dataset: The UT-Interaction dataset [29] contains 60 RGB video clips of 6 balanced interaction categories: *shake hands*, *hug*, *kick*, *point*, *punch* and *push*. Each class contains 10 video clips. The videos are challenging due to low resolution, background variations, and body part occlusions. Compared with the SBU dataset, the UT-Interaction contains fewer sequences in each category but longer duration with more frames in each captured sequence. We augment the dataset by mirroring the videos and halfway clipping. To extract skeleton positions, we use OpenPose [31] to detect the 2D postures of two characters in each image frame and collect the 15 main joints [17] in each character for our experiment.

B. Evaluations on the SBU Dataset

We first compare the performance of our methods against relevant algorithms tested on the SBU dataset including SVM-based classification on raw skeleton [17], interaction mesh-based motion features [5], sequential-based learning using RNN or its alternative LSTM [7]–[9], [22], and CNN on skeletal images [15], [24]. From the recognition accuracy in Table I, the single-stream PJD-BiLSTM achieves better performance using interaction descriptors compared with most of the baselines modeling single characters, and FCGC-BiLSTM further improves the classification result by mining the spatial correlations of joints through an adaptive graph, which already achieves the state of the arts. Finally, PJD+FCGC shows the best among all the comparisons, and we also observe that there is about 2% significant performance improvement in PJD+FCGC against the individual streams, which highlights the strength of our late fusion algorithm.



	approach	depart	kick	push	shake hands	hug	exchange	punch
approach	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
depart	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
kick	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
push	0.00	0.00	0.00	0.93	0.00	0.02	0.00	0.05
shake hands	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.17
hug	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
exchange	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.03
punch	0.05	0.00	0.00	0.16	0.02	0.00	0.02	0.74

(a) PJD-BiLSTM

	approach	depart	kick	push	shake hands	hug	exchange	punch
approach	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
depart	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
kick	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.05
push	0.00	0.02	0.00	0.95	0.00	0.00	0.00	0.02
shake hands	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.17
hug	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
exchange	0.00	0.00	0.00	0.00	0.05	0.00	0.93	0.02
punch	0.00	0.00	0.00	0.08	0.00	0.00	0.02	0.90

(b) FCGC-BiLSTM

	approach	depart	kick	push	shake hands	hug	exchange	punch
approach	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
depart	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
kick	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
push	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.02
shake hands	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.17
hug	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
exchange	0.00	0.00	0.00	0.00	0.02	0.00	0.98	0.00
punch	0.00	0.00	0.00	0.05	0.02	0.00	0.02	0.90

(c) PJD+FCGC

Fig. 3. Confusion matrices of separate streams and the late fusion on the SBU dataset.

TABLE II Recognition performance on the UT-Interaction dataset under different feature modalities.

Modality	Method	Acc.(%)
RGB	DBoW [32]	85.0
	MSSC [33]	83.3
	HR [34]	88.4
	IP [35]	91.6
	PKM [36]	93.3
RGB+skeleton	PA-DRL [37]	**96.7**
skeleton	PJD-BiLSTM	91.9
	FCGC-BiLSTM	92.7
	PJD+FCGC	94.4

To further investigate the contributions of individual streams and the improvement of late fusion, we present the three confusion matrices for different interaction classes in Fig. 3. Comparing the two streams, FCGC-BiLSTM outperforms PJD-BiLSTM on average, and we further observe that FCGC is better at discriminating between very similar interactions such as *push* and *punch*, but less effective in actions like *kick* and *exchange*. This is because PJD-BiLSTM can better tell the relative distance change. For example, the distances between the active leg with other joints can better describe *kicking* than positional features. While PJD-BiLSTM is less effective when discriminating similar interactions such as *punch* and *push* because their corresponding PJDs are also similar. When using FCGC-BiLSTM, the adaptive graph can detect the discriminative joint correlations to better classify these two interactions. This justifies the use of the two streams to complement each other in handling different types of interactions.

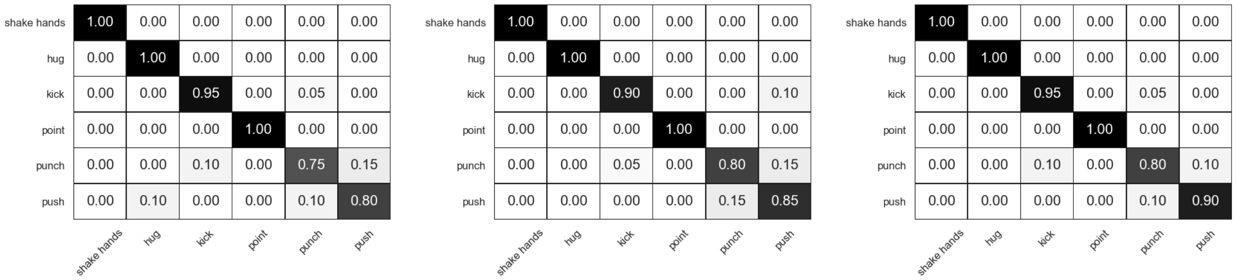
In Fig. 3c, the fusion accuracies are higher than (i.e., *push* and *exchange*) or at least equal to the best classification result between PJD-BiLSTM and FCGC-BiLSTM, which highlights the effectiveness of further improving the classification performance by combining the output of the two streams. The proposed late fusion alone does not introduce extra parameters to learn, and it is expected to be easily extended to other recognition work for performance boosting when multiple classifiers in conjunction with features in arbitrary size and length.

C. Evaluations on the UT-Interaction Dataset

Since the principle behind the network is using the pairwise distance among joint positions, our method can be reasonably adapted to the 2D pose domain. Inspired by this, we further challenge our framework on the UT-Interaction dataset solely based on skeletons and compare with the state-of-the-art RGB-based or RGB+skeleton approaches. Here, we only change the input data from 3D to 2D without altering the proposed network architecture. We have achieved a comparable performance as shown in Table II. The result of our PJD+FCGC outperforms the models using RGB features only, which justifies the effectiveness of modeling interactions using skeleton joint features over color information, as the latter can be largely affected by the unnecessary background color and the appearance of clothes. Our skeleton-based approach is less effective than PA-DRL [37] which combines both RGB and skeletal features, and we believe that their RGB features that describing the body part pixel changes can be integrated into our system to further improve the recognition performance. Within our two-stream method on this 2D dataset, we discover a consistent observation as in the experiment of the SBU dataset: FCGC-BiLSTM boosts the accuracy with the design of graph convolution on body joints compared with PJD-BiLSTM, and the late fusion of PJD+FCGC performs better than the individual streams.

We also present the result of three confusion matrices on the UT-interaction in Fig. 4. Similar to Fig. 3, the two streams perform differently on *kick*, *punch*, and *push* interactions. The fusion algorithm achieves the highest accuracies for all classes, and the result on *push* also proves that the fusion outperforms both streams.

We further investigate the possible reasons that may cause wrong predictions in some cases. We found that OpenPose suffers from information loss for the key joints when extracting the joint positions, which commonly appears in *punch*, *hug*, and *kick*. It fails to detect some limbs because of the partial occlusions of some body parts, such as the bottom row of the *punch* and *kick* interactions in Fig. 5. This leads to an ill impact on our approach since FCGC-BiLSTM largely depends on a reliable joint position input. While in most of the circumstances, our method could correctly predict the



(a) PJD-BiLSTM

(b) FCGC-BiLSTM

(c) PJD+FCGC

Fig. 4. Confusion matrices of separate streams and the late fusion on the UT-Interaction dataset.

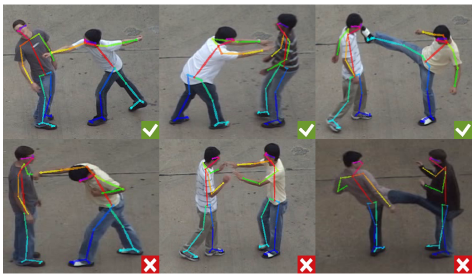


Fig. 5. Example frames with classification results in UT-Interaction dataset. From left to right columns, the ground truth labels are *punch*, *push*, and *kick*. The first row gives correctly classified samples, and the interactions in the second row are wrongly classified as *push*, *hug*, and *push*, respectively.

TABLE III Influence of different components in our proposed two-stream network.

Type	Method	JP	PJD	A	bi	Acc.(%)
Non graph	JP-BiLSTM	✓		N/A	✓	88.4
	PJD-LSTM		✓	N/A		91.4
	PJD-BiLSTM		✓	N/A	✓	**94.0**
Graph based	FCGC-BiLSTM w/o PJD+A	✓			✓	89.9
	FCGC-BiLSTM w/o PJD	✓		✓	✓	90.9
	FCGC-LSTM	✓	✓	✓		92.4
	FCGC-BiLSTM	✓	✓	✓	✓	**95.1**

interaction label if more joints are detected with a higher confidence rate.

D. Ablation Study

We conduct detailed ablation experiments on the SBU dataset to justify the effectiveness of using joint position (JP), pairwise joint distance (PJD), adaptive graph connectivity A, and the bi-directional encoding in our framework. The results are shown in Table III. As the baseline of our approach, we evaluate the first stream (non-graph models) by only encoding the joint positions without PJD (denoted as JP-BiLSTM). JP-BiLSTM has a much lower accuracy than PJD-BiLSTM, and this is because the generic RNN-based model is less effective in modeling the spatial interdependencies from the raw joint input compared with PJD which already contains spatial information among joints.

In the second stream (graph-based models), we conduct graph convolutions on joint positions to model the spatial correlations and incorporate it with PJD for its superior discriminative ability. Here, JP is a necessary component as each joint represents a basic graph node. We first perform plain graph convolutions on the raw positions without the dot product between PJD and the adaptive connection A in Eq. (3) (namely FCGC-BiLSTM w/o PJD+A). In this case, the graph connectivity follows the skeletal structure, which means that there will not be any joint connections between two characters, and only the connected joints in each character can affect each other. Compared with JP-BiLSTM, the recognition result increases (89.9% over 88.4%), which justifies the use of graph convolution. We then make A to be adaptive with the network training (namely FCGC-BiLSTM w/o PJD). We found that the accuracy further improves (90.9% over 89.9%) with more latent connections that not existed in the skeletal structure, such as the hand joint(s) of one character and the shoulder joint(s) of the other character that are closely correlated for differentiating *push* with the other interactions. FCGC-BiLSTM outperforms FCGC-BiLSTM w/o PJD by a large margin (95.1% over 90.9%), which indicates the effectiveness of incorporating PJD to represent the spatial proximity within all joints between two characters. Note that we do not perform the graph convolution between JP and PJD under fixed skeleton (FCGC-BiLSTM w/o A), as it only results in bone lengths which are irrelevant for classifying interactions.

From both PJD-LSTM vs. PJD-BiLSTM and FCGC-LSTM vs. FCGC-BiLSTM, we observe that the bidirectional propagation is essential in improving the performance, as it integrates both the past and future knowledge to solve the long term information loss problem. The two streams proposed in this work come from the best non-graph model PJD-BiLSTM and the best graph-based model FCGC-BiLSTM, and the accuracy can be further improved through fusing the two streams as shown in PJD+FCGC from Table I and Table II.

V. Conclusion

In this paper, we propose a two-stream architecture to classify human interactions by leveraging interactive-based geometric features. The two streams are 1) PJD-BiLSTM to encode the dynamics of pairwise distances among joints of the two characters with a BiLSTM network, 2) FCGC-BiLSTM to encode PJD into an adaptive graph convolution to learn the spatial proximity among joint nodes, and combining with BiLSTM to learn temporal transitions of underlying spatial proximity. We further propose a late fusion method based on the probabilistic model to generate an optimal score distribution for final prediction. The experimental results on both 3D and 2D benchmark datasets show the effectiveness and generality of the proposed model. In the future, we are interested in combining the skeleton position with color appearance to improve the recognition performance on RGB-based interaction datasets.

Acknowledgment

The project is supported in part by grants from City University of Hong Kong (Project No. 9220077 and 9678139), and the Royal Society (Ref: IES\R2\181024 and IES\R1\191147).

References

[1] H. P. Shum, T. Komura, M. Shiraishi, and S. Yamazaki, "Interaction patches for multi-character animation," *ACM Transactions on Graphics*, vol. 27, no. 5, pp. 1–8, 2008.

[2] E. S. L. Ho, T. Komura, and C.-L. Tai, "Spatial relationship preserving character motion adaptation," *ACM Transactions on Graphics*, vol. 29, no. 4, pp. 1–8, 2010.

[3] E. S. L. Ho and T. Komura, "Indexing and retrieving motions of characters in close contact," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 3, pp. 481–492, 2009.

[4] J. K. Tang, J. C. Chan, H. Leung, and T. Komura, "Interaction retrieval by spacetime proximity graphs," *Computer Graphics Forum*, vol. 31, pp. 745–754, 2012.

[5] E. S. L. Ho, J. C. P. Chan, Y.-m. Cheung, and P. C. Yuen, "Modeling spatial relations of human body parts for indexing and retrieving close character interactions," in *ACM Symposium on Virtual Reality Software and Technology*, 2015, pp. 187–190.

[6] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *IEEE Winter Conference on Applications of Computer Vision*, 2017, pp. 148–157.

[7] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1647–1656.

[8] Z. Fan, X. Zhao, T. Lin, and H. Su, "Attention-based multiview re-observation fusion network for skeletal action recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 363–374, 2018.

[9] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 499–508.

[10] J. K. T. Tang, H. Leung, T. Komura, and H. P. H. Shum, "Emulating human perception of motion similarity," *Computer Animation and Virtual Worlds*, vol. 19, no. 3–4, p. 211–221, 2008.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence*, 2018.

[13] K. Thakkar and P. Narayanan, "Part-based graph convolutional network for action recognition," in *British Machine Vision Conference*, 2018.

[14] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[15] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3288–3297.

[16] H. P. H. Shum, T. Komura, and S. Yamazaki, "Simulating competitive interactions using singly captured motions," in *ACM symposium on Virtual Reality Software and Technology*, 2007, pp. 65–72.

[17] 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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 28–35.

[18] Y. Shen, L. Yang, E. S. L. Ho, and H. P. H. Shum, "Interaction-based human activity comparison," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

[19] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *AAAI Conference on Artificial Intelligence*, 2016.

[20] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *IEEE International Conference on Computer Vision*, 2017, pp. 2117–2126.

[21] 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 *International Joint Conference on Artificial Intelligence*, 2018, pp. 786–792.

[22] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*, 2016, pp. 816–833.

[23] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *European Conference on Computer Vision*, 2018, pp. 103–118.

[24] Y. Tas and P. Koniusz, "Cnn-based action recognition and supervised domain adaptation on 3d body skeletons via kernel feature maps," in *British Machine Vision Conference*, 2018, pp. 1–13.

[25] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1227–1236.

[26] M. Li and H. Leung, "Graph-based approach for 3d human skeletal action recognition," *Pattern Recognition Letters*, vol. 87, pp. 195–202, 2017.

[27] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *IEEE International Conference on Computer Vision*, 2019, pp. 9489–9497.

[28] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 26–37, 1980.

[29] M. S. Ryoo and J. Aggarwal, "Ut-interaction dataset, icpr contest on semantic description of human activities (sdha)," in *IEEE International Conference on Pattern Recognition Workshops*, vol. 2, 2010, p. 4.

[30] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[31] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[32] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *International Conference on Computer Vision*, 2011, pp. 1036–1043.

[33] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. Mark Siskind, and S. Wang, "Recognize human activities from partially observed videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2658–2665.

[34] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *European Conference on Computer Vision*, 2014, pp. 689–704.

[35] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptions for human interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1775–1788, 2014.

[36] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2650–2657.

[37] L. Chen, J. Lu, Z. Song, and J. Zhou, "Part-activated deep reinforcement learning for action prediction," in *European Conference on Computer Vision*, 2018, pp. 421–436.