Computing and Informatics, Vol. 44, 2025, 408–428, doi: 10.31577/cai_2025_2_408

LIGHTWEIGHT DUAL-STREAM HUMAN BEHAVIOR INFERENCE NETWORK BASED ON MULTI-LAYER PERCEPTRON

Qichen Qin, Xiaohong Han*

Taiyuan University of Technology
College of Computer Science and Technology/College of Data Science
Taiyuan, 030000, China
e-mail: qinqichen@163.com, hanxiaohong@tyut.edu.cn

Abstract. The recognition of human behaviors in videos is a critical domain within human activity analysis. However, the current architectures and mechanisms of human behavior recognition methods dominated by CNN, GCNs, and LSTM are unduly complex resulting in high computational complexity of the models. Furthermore, these methods often exhibit poor robustness when it comes to recognizing behaviors across different environmental conditions and video angles. To address these challenges, this paper introduces a lightweight human skeleton interaction behavior inference network based on a multi-layer perceptron. This network leverages human skeleton information and utilizes minimal prior knowledge to infer limb behavior encoding. To reduce computational complexity, videos are divided into smaller segments, serving as the minimum computation units. This approach integrates three essential types of information: independent global information about individual postures, local interaction information regarding different limb parts, and temporal distance information. These three types of information are coupled through LSTM, incorporating temporal changes into network for recognition and classification. In comparison to previous similar methods, our proposed method is more lightweight, exhibits stronger robustness against interference and enables behavior recognition across different environments and perspectives.

Keywords: Human-human interaction, human action recognition, skeleton joints

* Corresponding author

1 INTRODUCTION

Recognizing human interaction behaviors in videos is a highly significant field of video comprehension with diverse applications. For example, in security surveillance, human behavior recognition technology can identify and track suspicious activities like intruders, thieves, or violent criminals. In healthcare, it can monitor and assess patients' movements, postures, physical conditions, and rehabilitation progress. In sports, it aids in analyzing athletes' movements and skills for performance enhancement.

Existing human behavior recognition technologies include models based on RGB videos and models based on skeleton sequences. Compared to RGB video-based methods, skeleton sequence-based models reduce errors related to individual appearance differences while lowering computational complexity, making them more attractive to researchers. These models can be categorized as traditional manual methods and neural network methods. Traditional manual methods offer interpretability but require intricate preprocessing. In contrast, neural network methods do not depend on prior knowledge but may not fully capture the semantic information of interactions between body parts. Considering the advantages and disadvantages of traditional manual methods and neural network methods, this paper adopts the traditional method's prior knowledge to guide the learning of neural networks, thereby achieving human behavior recognition.

Due to the simplicity of single skeletal information, many researchers have conducted multimodal experimental attempts [1, 2, 3, 4, 5, 6, 7], and have achieved certain results. Although more multimodal data can improve the accuracy of human behavior recognition, such methods require more parameters and training resources. Therefore, this study builds on multimodal concepts by transforming single modality into multiple branched modalities, simplifying data processing and enabling analysis from various perspectives.

Compared to single-person behavior recognition, dual-person behavior recognition not only considers the characterization of individual behavior postures but also models the complex interaction relationships of interaction positions in dual-person interactions. Unlike whole-body networks, the part-based neural networks require more sub-models to model each part. Cheng et al. [3] treats each keypoint as a part and uses the Relational Network [8] for relational inference to achieve recognition of human interactive behaviors. Ji et al. [9] divides the human body into five parts and generates eight limb interaction pairs to create an interaction dictionary for behavior recognition. The part-based neural network methods extract spatial and temporal features separately when extracting features from body parts and finally combine the two dimensions of features. Such methods fail to extract the spatio-temporal information implicit in motion information. To address these issues, this study considers the spatio-temporal relationship between the spatial and temporal dimensions when extracting motion information from each part. Semantic fusion is performed in the initial stage of feature extraction in both dimensions, resulting in more representative motion features for accurate human behavior recognition.

Given the shortcomings of the previous work, this paper proposes a novel, simple, and efficient skeleton-based lightweight dual-person interaction behavior recognition network. The main contributions are as follows:

1. This paper proposes a simple and lightweight dual-stream network to address the issue of high computational complexity in previous methods.
2. This paper proposes a semantic module based on interactive positions, which can extract features from different interactive positions separately, thus obtaining semantic descriptions of different interactive positions.
3. This paper proposes a single-person pose spatial feature representation method based on polar coordinates, which has better resistance to interference and robustness compared to previous methods.

The proposed network architecture has been experimentally validated on two datasets, demonstrating the effectiveness of the proposed approach.

The organization of the rest of the paper is as follows: In Section 2, we introduce the related research to this work. In Section 3, we discuss the overall model. The experimental results are detailed in Section 4. Finally, we present the conclusions of this work in Section 5.

2 RELATED WORK

2.1 Skeleton-Based Recognition Network

Existing techniques for human activity recognition primarily include recognition models based on RGB video and recognition models based on skeleton sequences. Compared to recognition methods based on RGB video, the skeleton-based recognition models are capable of reducing errors caused by individual differences in appearance, while also lowering computational complexity. As a result, this approach has garnered greater attention from researchers [10, 11, 12, 13, 14]. Network models based on skeleton sequences primarily fall into two categories: traditional manual approaches and neural network approaches. Traditional manual approaches involve extracting features from skeleton information based on prior knowledge. These methods can subjectively describe the characteristics of human behavior, thus achieving better results in behavior recognition. Based on the principle that "better view leads to better recognition", a skeleton-based HDS-SP descriptor is proposed for human activity recognition [15]. Although such a method offers excellent interpretability, it requires complex manual inference and design. On the other hand, neural network methods focus on data-driven learning of skeleton information, allowing for modeling of human activity patterns and achieving accurate human activity recognition. Yan et al. [16] proposes an extension of graph neural networks to spatiotemporal graph models (ST-GCN) to integrate information in both temporal and spatial dimensions, thereby achieving human behavior recognition. Building upon the foundation of ST-GCN, the traditional convolutional

operators are replaced by Shift convolutional operators [17] in order to enable the network model to achieve better performance with fewer parameters and computational resources [3]. Additionally, a novel decoupled spatiotemporal attention network (DSTA-Net) [18] is introduced, emphasizing the distinctive features of different motion scales to achieve accurate human behavior recognition. Although the neural network methods mentioned above have achieved satisfactory results, they tend to overlook local motion-type features, thereby resulting in incomplete learning of input skeletal sequences. To reconcile the advantages and disadvantages of both traditional manual and neural network methods, this paper integrates prior knowledge from the traditional approach into the neural network learning process. This approach enhances the ability of neural networks to accurately capture the implicit mappings in human behavior interactions, ultimately improving human behavior recognition.

2.2 Multimodal Behavior Recognition Network

Early multimodal human behavior recognition networks predominantly centered on modeling and predicting the physical attributes of skeletons [3, 4]. Recent studies have concentrated on elastic modeling of the topological structure at the channel level to achieve improved results [6, 7]. Some researchers have combined the physical attributes of skeletons and the RGB features of local image patches to focus on modeling crucial interactive regions. By concatenating these features with whole-body characteristics, they have successfully implemented human behavior recognition [2]. An approach has been proposed to unify the processing of various modalities [5]. This approach's network architecture can dynamically adapt based on different data modalities or node quantities, facilitating interaction fusion among modalities. Duan et al. [1] introduced a novel modality known as 3D heatmaps, generating heatmaps for each frame by employing Gaussian kernels on skeletal point coordinates. These heatmaps are then layered for the entire sequence, enabling the recognition of human interactive behaviors. While multimodal data can enhance network recognition performance, it introduces complexity to the network structure and increases the number of parameters. Consequently, this study adopts a single modality processed as multiple branch modalities, ensuring simplicity in data processing while enabling analysis and semantic representation of the same modality from different perspectives.

2.3 Part-Based Recognition Network

Part-based networks prioritize the interactive parts involved in human interactive behaviors. These methods model each individual part to acquire semantic information about the interactions between these parts in human interactive behaviors. Perez et al. [11] treat each joint as a body part and individually model the relationships between these parts. They use the Relational Network [8] for inferring

interactive relationships, resulting in the recognition of human interactive behaviors. Ji et al. [9] divided the human body into five parts and generated eight limb interaction pairs. They employed a contrast mining algorithm to identify significant interaction pairs within each interaction category, creating an interaction dictionary for behavior recognition. Lee and Lee [2] introduced an attention mechanism among interacting body parts, focusing selectively on local movements between body joints. Additionally, body behavior recognition is achieved through the modeling of the entire body and the co-occurrence matrix of sub-volumes. While these methods have enhanced the accuracy of human behavior recognition, they tend to overlook the spatial and temporal connections in the extraction of bodily features, as well as the implicit spatiotemporal information in motion behavior, hindering human behavior recognition. To address this, our study takes into account the spatiotemporal features inherent in both spatial and temporal dimensions. Semantic fusion is conducted at the initial stage of feature extraction, aiming to obtain more representative motion features for the precise recognition of human behavior.

3 METHOD

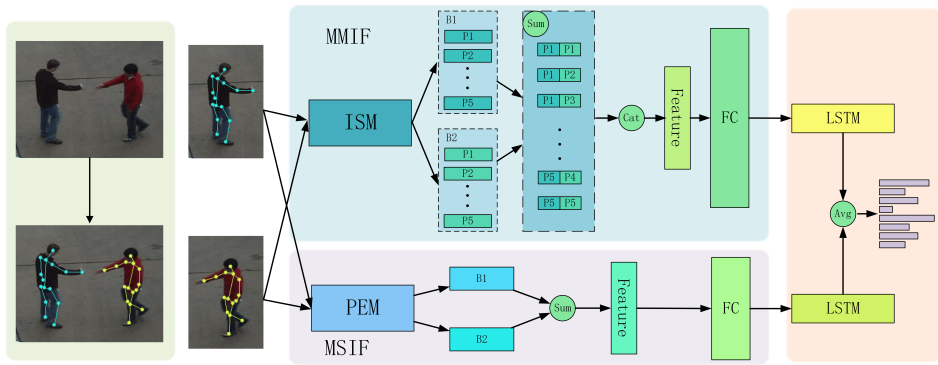

Figure 1. Overview of the proposed framework. The network architecture includes two data flows, spatial and motion flow. The dashed box contains the PSM processing unit which executes motion flow processing.

In this paper, we provide a detailed explanation of the network architecture proposed for human behavior recognition. The proposed network structure offers a more straightforward principle and a clearer framework than previous architectures. Its effectiveness has been validated using popular human behavior recognition datasets. The overall structure of the framework is illustrated in Figure 1, which consists two sub-networks: the primary motion information flow and the primary spatial information flow.

Given two interacted skeletal sequences $B_l, B_r \in R^{T\times J\times 2}$, where T and J represent the number of frames and the number of nodes, respectively, and 2 rep-

resents the x and y dimensions of each node coordinate. In the primary motion information flow, we initially feed the skeletal sequences into the Initial Semanticization Module (ISM) to process the original skeletal sequences and obtain higher-dimensional primary semantic features. We then model the individual body parts through the Part Semanticization Module (PSM) and perform inference using the Limb Part Interaction Reasoning Module to obtain the final classification features of this flow network. In the main spatial information flow, the skeletal sequence is the first input to the Pose Encoding Module (PEM) for encoding spatial information. Subsequently, the pose features of the interacting parties are fused to obtain classification features. The classification features from the two information flows are then obtained by using LSTM for temporal classification. The two classification results are averaged to generate the final recognition results.

3.1 Main Motion Information Flow (MMIF)

In this section, we provide an overview of the primary motion information flow based on skeleton joints. Firstly, we introduce the initial semantic module, followed by an exploration of the processing methods employed in the motion information flow of the part semantic module (PSM). We then delve into the module for inter-inference between body parts.

3.1.1 Initial Semantic Module (ISM)

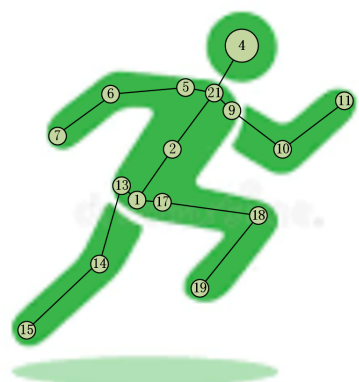

Figure 2. The skeletal sequence is partitioned into five interacting body parts (torso, left arm, right arm, left leg, and right leg), with each part corresponding to specific index positions of the skeleton joints

Directly using the original skeletal sequences B_l, B_r for network learning and modeling is computationally demanding and often yields unsatisfactory outcomes. Empirical findings suggest that incorporating manual processing can enhance network learning, akin to the ResNet structure [4], where manual mappings compensate for inadequate network learning. In this module, the residual connection is established by linking the original skeletal sequences B_l, B_r with the processed initial semantic features I_l, I_r through the following equation:

$$I_l' = FC\left(ISM\left(B_l\right)\right) + B_l, \quad I_r' = FC\left(ISM\left(B_r\right)\right) + B_r,$$

where FC represents the fully connected layer, which performs dimensionality augmentation on the I_l and I_r generated by ISM, and combines them with B_l and B_r through summation.

Experimental results reveal that the residual connection provides limited accuracy improvement while increasing network complexity and parameters. Thus, this study solely employs the manually guided ISM output as the input for the subsequent stage.

For a given pair of interacting skeletal sequences B_l, B_r, all joints are initially partitioned into five primary parts (right arm P_1, left arm P_2, right leg P_3, left leg P_4, torso P_5). Each subset contains an equal number of joints, and the entire video is divided into N segments to reduce computational complexity. The primary semantic analysis of each video segment follows these steps:

Velocity v**:** Calculate the velocities of the endpoint and connection point for each subpart P_i of every interactive individual. Speed is determined by analyzing the displacement change between frames, and the average velocity of each video frame is used as the speed value for that segment.

$$v_j^i = \frac{\sqrt{\left(x_j^{T+t} - x_j^T\right)^2 + \left(y_j^{T+t} - y_j^T\right)^2}}{t},$$

where i, j represent the encoding of body parts and the index of joint nodes, respectively, while x, y represent the coordinates of the two dimensions. T, t represent the starting time of each video segment and the duration of each segmented sub-video, respectively.

Acceleration a**:** Given small time dimensions in segmented sub-video segments, the motion is considered uniform acceleration. Acceleration a is calculated by analyzing the displacement difference within equal time intervals. The changes in velocity and acceleration between different segments clearly indicate the level of involvement of each body part in the interactive process.

$$a_j^i = \frac{\Delta \mathrm{s}}{t^2},$$

where i, j represent the encoding of body parts and the index of joint nodes,

respectively. Δs represents the displacement difference in equal time frames, and t represents the duration of each segmented sub-video.

Distance dist: Dynamic characteristics of proximity and separation between interacting body parts can be modeled through distances between these parts. In the course of interaction, the interacting body parts are typically the extremities of the limbs, such as the hands and feet. Therefore, for participant B_l, the distance characteristics of their interaction are represented by calculating the shortest distance between the endpoints of each body part and all the nodes of the corresponding body parts of the other participant B_r.

$$\mathrm{dist}_{rk}^{i} = \min\left(\mathrm{Dist}\left(\mathrm{Joint}^{P_i^l}, \mathrm{Joint}_1^{P_k^r}\right), \ldots, \mathrm{Dist}\left(\mathrm{Joint}^{P_i^l}, \mathrm{Joint}_n^{P_k^r}\right)\right),$$

where i, k represent the encoding of the body parts. l, r represent the indexes of the interacting participants. n represents the index of joint node j. Dist() is the function used to calculate the Euclidean distance between two joints.

Joint angle θ: Joint angles effectively represent spatial information in behavior recognition and play a supplementary role in the motion information flow.

$$\theta = \arccos\left(\frac{l_1 \cdot l_2}{\|l_1\| \cdot \|l_1\|}\right),$$

where l_1, l_2 respectively represent the two lines on either side of the angle being calculated.

The behavioral information of each body part of the interacting individuals includes joint velocity, joint angle, and joint acceleration. These features possess translational invariance properties similar to convolutional networks, effectively reducing data volume while maintaining integrity.

3.1.2 Part Semantic Module (PSM)

The Part Semantics Module (PSM) aims to provide a semantic description of the motion behavior for each interacting body part in interaction activities. A simple Multi-Layer Perceptron (MLP) model is employed to capture the behavior of the body parts. While fully connected networks require more parameters and computations, they make efficient use of computer resources, leverage temporal and spatial locality, and accelerate network computation. Additionally, well-designed fully connected networks, as shown by RNs [8], exhibit strong reasoning capabilities. In this paper, an MLP is used to semantically analyze the motion information of interacting body parts.

As shown in Figure 3, the PSM model for modeling individual body parts comprises only four very narrow fully connected layers. The first two layers map limb motion information to a high-dimensional feature space, while the last two layers infer the mapped features and introduce residual structures [19] to enhance the model's training speed.

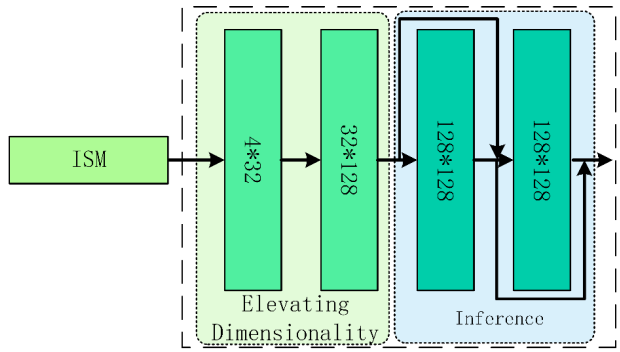

Figure 3. The overall presentation of the Part Semantic Module is as follows: the input is the semanticized results of the ISM, which undergo dimensionality expansion and reasoning processes to obtain the final interactive semantic features

3.1.3 Limb Interacting Inference Module (LIIM)

In interpersonal interactions, the mutual movements of body parts of both individuals play a vital role in behavior recognition. Recognizing dual-person behavior involves considering not only individual posture representation but also modeling complex interaction relationships between the two individuals. For instance, in the act of patting someone, it is challenging to determine whether one person's hand extension targets the waist, hand, head, shoulder, or other parts of the other person, as it involves multiple interaction analogies. Given the aforementioned observations, it is natural to consider that in dual person behavior recognition, there exists an inherent correspondence between the interacting body parts of the two individuals, and their movements exhibit potential complementarity and imply semantic relevance.

For example, in a handshake action, the right hands of both individuals extend sequentially to make contact and then separate. Exploring the semantic correlation between the interacting body parts can effectively contribute to understanding dual person behavior interactions in videos. Additionally, for certain interaction behaviors, the variations in distance between the interacting body parts also present distinct and discernible features. For instance, in hugging activities, the trunk parts of both individuals gradually move closer over time until they make contact, sustaining that contact for a period before separating again.

To capture these aspects, we divide the human body into parts, such as the torso, left arm, right arm, left leg, and right leg. Each part of an individual can interact with the corresponding part of another person or with other parts of the other person. Interactions between the same parts are the most common and have higher weight, while interactions between different parts are less frequent but still significant. Self-adjustment of weights is applied to the intensity of part interactions.

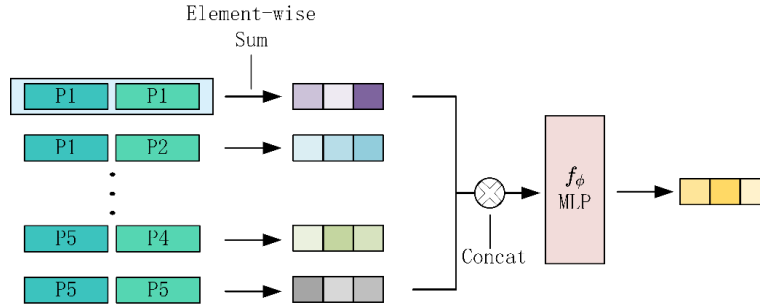

Figure 4. The overall architecture of the Limb Interacting Inference Module (LIIM). The individual parts of the interacting parties are fused at the elemental dimension and information coupling is achieved through a Multilayer Perceptron (MLP).

The sequential relationship between interacting parties is crucial in interactive behaviors, and different interaction orders affect the model's internal parameters. To address this, we use Relational Networks [8] as the underlying network, employing the method of adding feature vectors. While this increases the solution space of feature vectors and reduces reliability somewhat, it eliminates the need to consider interaction order and simplifies determining the initiator and recipient of the interaction.

After the semantic representation of the movement features of each body part through PSM, interactive pairs are formed, and the features of interacting parts within each pair are added together to obtain the feature vector specific to each interaction pair. The twenty-five pairs of interactive relationships generated by the five body parts are concatenated to obtain the interaction relation feature vector. Further relationship inference is conducted through the MLP layer to obtain the dual-person interactive motion feature vector of the main movement information flow. The above content can be formalized as follows:

$$\text{Feature}_{\text{out}} = f_\phi g_{\text{concat}}(P_1^l + P_1^r), (P_1^l + P_2^r), \ldots, (P_5^l + P_4^r), (P_5^l + P_5^r),$$

where $\text{Feature}_{\text{out}}$ represents the output features of the inference module. l and r represent the indices of the interacting parties. P_i^l represents the feature vector of the i^{th} body part of the party at index l. f_ϕ denotes the inference function of the body module, and g_{concat} denotes the feature concatenation module for interaction.

The approach presented in this paper, when coupled with sufficient data and appropriate parameter settings, achieves human activity recognition performance comparable to the Bag-of-Words (BoW) model, all without requiring manual clustering or similar preprocessing.

3.2 Main Spatial Information Flow (MSIF)

In this section, we introduce the main spatial information flow based on skeleton joints. The first subsection of this chapter explores the spatial characteristics of an individual in interaction behavior and employs polar coordinates for pose abstraction. Subsequently, we extend the previously mentioned PSM module to model the inference of individual poses.

3.2.1 Posture Encoding Module (PEM)

The posture encoding module aims to encode the spatial appearance and pose of individual actors involved in interaction behavior. It is known from subjective experience that clearer judgments about the interaction category can be made when the independent actions of both parties in the interaction are discernible. For instance, in a handshake, both parties extend their right hand forward, and in a kicking action, one party extends their right leg forward and upward, while the other party may step back, among other possibilities. In dual-person interactions, the personal pose information of the interacting parties complements the interaction information between body parts, enhancing recognition accuracy. This paper utilizes a polar coordinate representation to describe the spatial appearance and pose of individuals and employs an expanded PSM module to encode and process the overall pose of an individual.

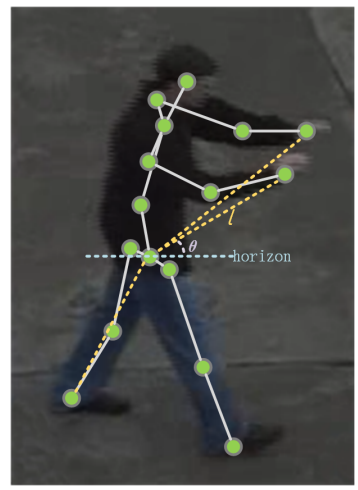

Figure 5. PEM module: single-person spatial posture feature. The features are visually displayed using the coordinates of the hands.

As depicted in Figure 5, the Pose Encoding Module (PEM) initiates a polar coordinate transformation on the individual's original skeletal information, with the

hip joint node as the coordinate origin, ensuring better feature stability in the extracted data. The processed data is subsequently fed into the expanded Pose-Specific Module (PSM) for encoding processing to obtain the final feature representation.

Additionally, the proposed module employs the feature addition method [8] to address sequencing issues between the interacting parties. This approach avoids redundant input data and approximates the effectiveness of a co-occurrence matrix [2], thus sidestepping coefficient-related problems. By adopting this approach, the objective becomes clearer, data duplication is reduced, and the network can reconstruct individual pose information. The formalization of this content is as follows:

$$\text{Feature} = \text{FC}\left(\text{PEM}\left(B_l\right) + \text{PEM}\left(B_r\right)\right),$$

where Feature represents the high-level feature that has been semantically interpreted by PEM, FC denotes the fully connected layer, and B_l, B_r represent the raw skeleton data of the interacting parties.

4 EXPERIMENTS

4.1 Datasets

NTURGB + D 60 datasets [14]: This dataset is a large-scale action recognition dataset that contains skeleton sequences. It comprises 40 different performers and includes 60 activity categories, with a total of 56 578 samples. The dataset was captured using three Microsoft Kinect v.2 cameras simultaneously. Each human skeleton in the dataset consists of 25 skeletal joints represented by 3D coordinates. The authors of the dataset propose two evaluation protocols:

1. Cross-Subject protocol, which divides the data based on subject ID, with samples from 20 subjects for training and the remaining subjects for testing; and
2. Cross-View protocol, which divides the data based on camera views, with samples from two camera views for training and the rest for testing.

The dataset includes a total of 11 human interaction categories, such as back slap, hand over, walking toward, kicking, pointing, touching the pocket, walking, pushing, punching, hugging, and shaking hands. The maximum number of frames in each sample is 256.

NTURGB + D 120 datasets [13]: This dataset is an expanded version of the NTU-RGB+D dataset, comprising an additional 60 action categories and 57 367 samples. It includes 113 945 skeleton sequences from 120 action categories and offers two standard evaluation protocols. The first protocol contains a subject division (Cross-Subject), where 53 subjects' actions are used for training, and the remaining subjects' actions are used for testing. The second protocol is called setup division (Cross-Setup), in which half of the participants are employed for

training and the other half for testing. Along with the 11 interaction categories in the NTU-RGB + D dataset, this dataset includes an additional 15 human interaction categories, bringing the total to 26 categories.

4.2 Implementation and Optimization Details

The method proposed in this paper is illustrated in Figure 1. The main motion information flow and the main spatial information flow are computed in parallel, and data inference is performed using a Multi-Layer Perceptron (MLP). The entire network is trained with random initialization using the PyTorch deep learning framework. The DINet architecture is implemented, with the ADAM optimizer and a base learning rate of 0.001. The weight decay rate is set to 0.001, and the number of epochs is set to be 3 000. The batch size is set to the default value of 512, and the cross-entropy loss function is used as the default loss function for the network. All models are trained on a server system equipped with a 12 GB Tesla K80 GPU.

During the data preprocessing stage, we address the missing keypoint information in frames by applying linear interpolation when only a small portion of the keypoints is missing. Frames with more than 50 % missing keypoints are discarded. To reduce computational load, we divide the videos into 10 segments, using them as metadata. As our proposed method requires segment-wise computation of physical attributes of keypoints, we replicate frames for samples with fewer than 40 frames, ensuring the minimum computational requirement. The specific implementation details of the proposed network architecture are shown in the following table. For convenience, activation functions and normalization functions are omitted in the table.

<table>
<tr><th>MMIF</th><th>MSIF</th></tr>
<tr><td>ISM</td><td rowspan="2">PEM</td></tr>
<tr><td>PSM</td></tr>
<tr><td>MLP$(25 * 128, 2\,048)$</td><td rowspan="2">MLP$(25 * 128, 2\,048)$</td></tr>
<tr><td>MLP$(25 * 128, 2\,048)$</td></tr>
<tr><td>MLP$(25 * 128, 2\,048)$</td><td>MLP$(25 * 128, 2\,048)$</td></tr>
<tr><td colspan="2">Flatten</td></tr>
<tr><td colspan="2">FC$(1\,280, 256)$</td></tr>
<tr><td>Drop(0.1)</td><td>Drop(0.2)</td></tr>
<tr><td colspan="2">FC(256, num_class)</td></tr>
</table>

Table 1. The specific details of the network architecture are illustrated in the diagram, which includes the three proposed modules, namely ISM, PSM, and PEM. The flatten layer is used for dimension expansion, and num_class denotes the number of categories in the dataset.

4.3 Ablations

The Impact of PEM: The PEM module is utilized to encode spatial information of individual behaviors. We conducted ablation experiments on this module using both single spatial information flow and spatiotemporal dual information flow. For each of the two validation information flows, we performed three comparative experiments:

1. Only Projection: the original data was directly projected onto the input dimension of the subsequent network for interaction behavior classification;
2. PEM + Res: the PEM approach was applied, and the output of PEM was connected with the original data through a residual connection;
3. PEM: the output of PEM was used as input for subsequent network to recognize interaction behaviors. Experimental results are presented in the table below.

In both MSIF and MSIF+MMIF, the network performance using only projection is the lowest among the two datasets. However, the network architecture of PEM+Res significantly improves the accuracy of human behavior recognition for both datasets. In the single MSIF experiments, compared to PEM + Res, using only PEM improves the accuracy by 5.8 percentage points on the NTURGB 60 dataset, while it decreases by 0.6 percentage points on the NTURGB 120 dataset. In the MSIF + MMIF experiments, the single PEM network structure outperforms the PEM + Res network by 1.4 percentage points on the NTURGB-D 60 dataset and 3 percentage points on the NTURGB-D 120 dataset.

	MSIF		MSIF + MMIF	
Data Set	NTU-D 60	NTU-D 120	NTU-D 60	NTU-D 120
Experiment	Acc		Acc	
only Projection	65.9	65.7	84.6	71
PEM + Res	68.5	70.1	87.3	73.7
PEM (ours)	74.3	69.5	88.7	76.7

Table 2. Comparison of PEM experiments: Acc represents the recognition accuracy of various methods on the dataset

The Impact of ISM: ISM is utilized to convert coordinate information of a specific body part into higher-dimensional patterns of motion and spatial information. To verify the effectiveness of ISM, we employed three different network structures, namely only Projection, ISM + Res, and DINet, on two datasets.

1. Only Projection directly projects the original data onto the input dimensions of the subsequent network.
2. ISM + Res applies ISM to transform the original coordinate information and connects the output of ISM with the original data through a residual connection.

3. DINet solely utilizes the output of ISM as input for the subsequent network. Experimental results are presented in the table below.

On the two datasets, the network using only Projection achieves accuracy rates of 82.7 % and 67.8 %, respectively. Compared to the single projection approach, the ISM+Res network significantly improves the accuracy of behavior recognition on both datasets. In comparison to the ISM + Res network proposed in this paper, DINet achieves an improvement of 3.1 percentage points on the NTURGB 60 dataset and an increase of 5.4 percentage points on the NTURGB 120 dataset.

	NTU-D 60	NTU-D 120
Experiment	Acc	Acc
ISM Projection	82.7	67.8
ISM + Res	85.6	71.3
without PSM	76.5	62.3
DINet	88.7	76.7

Table 3. Comparison of ISM and PSM experiments: Acc represents the recognition accuracy of various methods on the dataset

The Impact of PSM: PSM plays the role of semanticizing the motion information of body parts and acts as a higher-dimensional information processor. To evaluate the influence of PSM, we experimented with two different network structures, namely without PSM and DINet, on two datasets.

1. Without PSM: the PSM module is removed from the network, and the output features of ISM are directly fed into the subsequent network.
2. DINet: the standard network structure proposed in this paper, which includes PSM.

As shown in Table 3, on the NTURGB 60 dataset, the network model with PSM exhibits an improvement of 12.2 percentage points compared to the network structure without PSM. On the NTURGB 120 dataset, there is an improvement of 14.4 percentage points. These experimental results indicate that the proposed PSM achieves better performance on larger datasets, suggesting that the inclusion of PSM in the network effectively enhances the dimensionality and semantic interpretation of the input information, thus enhancing the recognition capability of the model.

The Impact of Branch Flows: To investigate the impact of two branch data flows on the experimental results, we conducted three different control experiments: MMIF, MSIF, and MMIF + MSIF on two datasets. Specifically, MMIF refers to the isolated main motion information flow, MSIF represents the isolated main spatial information flow, and MMIF + MSIF denotes the integration of both information flows. As shown in Table 4, the results indicate that in both datasets, the main motion information flow within a single branch outperforms the main spatial information flow. This aligns with the notion that

motion information plays a more crucial role compared to spatial information in dyadic interactive behaviors [20]. On the other hand, the fusion of motion and spatial information flows yields a better recognition of human interactive behaviors, demonstrating superior performance compared to all other ablation experiments and a notable enhancement over the single branch flow.

In summary, an overall analysis of the above ablation experiments reveals that the addition of PSM significantly improves the accuracy of recognition within the main motion information flow. This finding suggests that PSM plays a highly positive role in modeling the interactions between body parts, effectively implicitly mapping the patterns between these parts. In contrast, the impact of PEM within the main spatial information flow is not as pronounced as PSM. One possible reason for this difference is that although the polar coordinates used by PEM differ from Cartesian coordinates, they are merely alternative representations of the same data. The improvement in performance may result from manually converting Cartesian coordinates into polar coordinates, which substitutes some of the network's learning process and explicitly maps the patterns of human interaction.

	NTU-D 60	NTU-D 120
Experiment	Acc	Acc
MMIF	86.1	71.5
MSIF	74.3	65.3
MMIF + MSIF	88.7	76.7

Table 4. The impact of branch flows on model recognition performance

4.4 Result

In this study, we compare the performance of DINet, the method proposed in this paper, with other methods on the NTU RGB-D 60 and NTU RGB-D 120 datasets, as shown in Table 5. The NTU RGB-D 60 dataset employs the XSub and XView validation methods, while the NTU RGB-D 120 dataset uses the XSub and XSet validation methods. We categorize existing methods for human behavior recognition into three different framework modes based on different design approaches, including methods based on CNNs, methods based on RNNs, and methods based on GCNs, among others. Given that GCN-based methods have garnered more attention and achieved better results in the current research landscape, they have overshadowed further exploration of other types of methods in the field of human behavior recognition. Therefore, this paper explores the application of multilayer perceptrons in this domain through experimental investigation.

Based on the data in Table 5, it is evident that DINet outperforms the other listed methods in terms of recognition performance on NTU-60 (XSet, XView). Specifically, in the XSet validation of the NTU-60 dataset, DINet achieves an accuracy of 88.7 %, surpassing the highest performance of other methods at 88.6 %.

Model	NTU-60 XSub	NTU-60 XView	NTU-120 XSub	NTU-120 XSet
Synthesized CNN [21]	80.0	87.2	–	–
3scale ResNet [22]	85.07	92.3	–	–
STA-LSTM [23]	73.4	81.2	–	–
VA-LSTM [24]	79.2	87.7	–	–
ST-GCN [25]	81.5	88.3	70.7	73.2
3s RA-GCN [26]	87.3	93.6	81.1	82.7
2s-AGCN [27]	88.5	95.1	82.5	84.2
GR-GCN [28]	87.5	94.3	–	–
PGCN-TCA [29]	88.0	93.6	–	–
CoAGCN* [30]	84.1	92.6	80.4	82.0
3SCNN [31]	88.6	93.7	–	–
ours	88.7	95.2	76.7	80.3

Table 5. Comparison of DINet with existing excellent methods on two datasets using two validation approaches

In the XView validation of the NTU-60 dataset, DINet achieves a performance of 95.2 %, also surpassing other methods. Despite the performance drop of DINet on the more challenging NTU-120 dataset, with differences of 5.7 percentage points and 3.9 percentage points compared to 2s-AGCN [27] in the XSub and XSet validations, respectively, the performance remains competitive. Additionally, DINet, proposed in this paper, demonstrates simpler network structure and implementation compared to other networks for human interaction behavior recognition. Benefiting from the proposed interaction-part reasoning module, the network also achieves competitive results in terms of accuracy.

5 CONCLUSION AND FUTURE WORK

This study proposes DINet, a simple and lightweight Multilayer Perceptron (MLP) based network for human interaction behavior recognition. DINet aims to address the issues of complex network structures and suboptimal recognition performance under different viewing angles encountered by existing approaches. DINet adopts a popular dual-stream network architecture, consisting of the main motion information stream and the main spatial information stream. Specifically, the main motion information stream applies semantic processing to the interacting body parts, whereas the main spatial information stream describes the overall appearance of the interaction between two individuals. The fusion of these two data streams using the late fusion technique enables the effective modeling of dyadic interaction behavior. The proposed method is evaluated on the NtuRGB-D 60 [14] and NtuRGB-D 120 [13] datasets, and experimental results demonstrate that DINet, the proposed model, achieves recognition performance comparable to existing approaches while exhibiting simplicity in network structure and low computational requirements.

In future research, our focus will shift towards the study of multi-person interaction behavior recognition. This area represents a natural progression of our current research direction, as understanding multi-person interactions can provide deeper insights into human social behavior. By analyzing various cues such as postures, movements, and facial expressions, we aim to uncover patterns, motivations, and influencing factors of human social interactions. This research will not only contribute to the field of human behavior recognition but also provide valuable insights for social science and psychology.

To achieve this goal, we plan to broaden the validation of our models by testing them on diverse datasets. This will help us assess the generalizability and robustness of our models across different scenarios and populations. Additionally, we aim to enhance the real-time analysis capabilities of our models, as this is crucial for practical applications such as surveillance and interactive systems. By improving the efficiency and speed of our models, we can ensure their effectiveness in real-world scenarios.

REFERENCES

[1] Duan, H.—Zhao, Y.—Chen, K.—Lin, D.—Dai, B.: Revisiting Skeleton-Based Action Recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2959–2968, doi: 10.1109/CVPR52688.2022.00298.

[2] Lee, D. G.—Lee, S. W.: Human Interaction Recognition Framework Based on Interacting Body Part Attention. Pattern Recognition, Vol. 128, 2022, Art. No. 108645, doi: 10.1016/j.patcog.2022.108645.

[3] Cheng, K.—Zhang, Y.—He, X.—Chen, W.—Cheng, J.—Lu, H.: Skeleton-Based Action Recognition with Shift Graph Convolutional Network. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 180–189, doi: 10.1109/CVPR42600.2020.00026.

[4] Shi, L.—Zhang, Y.—Cheng, J.—Lu, H.: Skeleton-Based Action Recognition with Directed Graph Neural Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7904–7913, doi: 10.1109/CVPR.2019.00810.

[5] Trivedi, N.—Sarvadevabhatla, R. K.: PSUMNet: Unified Modality Part Streams Are All You Need for Efficient Pose-Based Action Recognition. In: Karlinsky, L., Michaeli, T., Nishino, K. (Eds.): Computer Vision – ECCV 2022 Workshops. Springer, Cham, Lecture Notes in Computer Science, Vol. 13805, 2022, pp. 211–227, doi: 10.1007/978-3-031-25072-9_14.

[6] Chen, Y.—Zhang, Z.—Yuan, C.—Li, B.—Deng, Y.—Hu, W.: Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13339–13348, doi: 10.1109/ICCV48922.2021.01311.

[7] Xu, K.—Ye, F.—Zhong, Q.—Xie, D.: Topology-Aware Convolutional Neural Network for Efficient Skeleton-Based Action Recognition. Proceedings of the AAAI

Conference on Artificial Intelligence, Vol. 36, 2022, No. 3, pp. 2866–2874, doi: 10.1609/aaai.v36i3.20191.

[8] Santoro, A.—Raposo, D.—Barrett, D. G.—Malinowski, M.—Pascanu, R.—Battaglia, P.—Lillicrap, T.: A Simple Neural Network Module for Relational Reasoning. In: Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc., 2017, pp. 4967–4976, https://proceedings.neurips.cc/paper_files/paper/2017/file/e6acf4b0f69f6f6e60e9a815938aa1ff-Paper.pdf.

[9] Ji, Y.—Ye, G.—Cheng, H.: Interactive Body Part Contrast Mining for Human Interaction Recognition. 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2014, pp. 1–6, doi: 10.1109/ICMEW.2014.6890714.

[10] Pang, Y.—Ke, Q.—Rahmani, H.—Bailey, J.—Liu, J.: IGFormer: Interaction Graph Transformer for Skeleton-Based Human Interaction Recognition. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., Hassner, T. (Eds.): Computer Vision – ECCV 2022. Springer, Cham, Lecture Notes in Computer Science, Vol. 13685, 2022, pp. 605–622, doi: 10.1007/978-3-031-19806-9_35.

[11] Perez, M.—Liu, J.—Kot, A. C.: Interaction Relational Network for Mutual Action Recognition. IEEE Transactions on Multimedia, Vol. 24, 2022, pp. 366–376, doi: 10.1109/TMM.2021.3050642.

[12] Li, C.—Hou, Y.—Wang, P.—Li, W.: Joint Distance Maps Based Action Recognition with Convolutional Neural Networks. IEEE Signal Processing Letters, Vol. 24, 2017, No. 5, pp. 624–628, doi: 10.1109/LSP.2017.2678539.

[13] Liu, J.—Shahroudy, A.—Perez, M.—Wang, G.—Duan, L. Y.—Kot, A. C.: NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, 2020, No. 10, pp. 2684–2701, doi: 10.1109/TPAMI.2019.2916873.

[14] Shahroudy, A.—Liu, J.—Ng, T. T.—Wang, G.: NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1010–1019, doi: 10.1109/CVPR.2016.115.

[15] Liu, J.—Wang, Z.—Liu, H.: HDS-SP: A Novel Descriptor for Skeleton-Based Human Action Recognition. Neurocomputing, Vol. 385, 2020, pp. 22–32, doi: 10.1016/j.neucom.2019.11.048.

[16] Yan, S.—Xiong, Y.—Lin, D.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018, No. 1, doi: 10.1609/aaai.v32i1.12328.

[17] Wu, B.—Wan, A.—Yue, X.—Jin, P.—Zhao, S.—Golmant, N.—Gholaminejad, A.—Gonzalez, J.—Keutzer, K.: Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9127–9135, doi: 10.1109/CVPR.2018.00951.

[18] Shi, L.—Zhang, Y.—Cheng, J.—Lu, H.: Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action-Gesture Recognition. In: Ishikawa, H., Liu, C. L.,

Pajdla, T., Shi, J. (Eds.): Computer Vision – ACCV 2020. Springer, Cham, Lecture Notes in Computer Science, Vol. 12626, 2021, pp. 38–53, doi: 10.1007/978-3-030-69541-5_3.

[19] He, K.—Zhang, X.—Ren, S.—Sun, J.: Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[20] Simonyan, K.—Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. Q. (Eds.): Advances in Neural Information Processing Systems 27 (NIPS 2014). Curran Associates, Inc., 2014, pp. 568–576, https://proceedings.neurips.cc/paper_files/paper/2014/file/ca007296a63f7d1721a2399d56363022-Paper.pdf.

[21] Liu, M.—Liu, H.—Chen, C.: Enhanced Skeleton Visualization for View Invariant Human Action Recognition. Pattern Recognition, Vol. 68, 2017, pp. 346–362, doi: 10.1016/j.patcog.2017.02.030.

[22] Li, B.—Dai, Y.—Cheng, X.—Chen, H.—Lin, Y.—He, M.: Skeleton Based Action Recognition Using Translation-Scale Invariant Image Mapping and Multi-Scale Deep CNN. 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2017, pp. 601–604, doi: 10.1109/ICMEW.2017.8026282.

[23] Song, S.—Lan, C.—Xing, J.—Zeng, W.—Liu, J.: An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, 2017, No. 1, doi: 10.1609/aaai.v31i1.11212.

[24] Zhang, P.—Lan, C.—Xing, J.—Zeng, W.—Xue, J.—Zheng, N.: View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2136–2145, doi: 10.1109/ICCV.2017.233.

[25] Zheng, S.—Lu, J.—Zhao, H.—Zhu, X.—Luo, Z.—Wang, Y.—Fu, Y.—Feng, J.—Xiang, T.—Torr, P. H. S.—Zhang, L.: Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6877–6886, doi: 10.1109/CVPR46437.2021.00681.

[26] Song, Y. F.—Zhang, Z.—Shan, C.—Wang, L.: Richly Activated Graph Convolutional Network for Robust Skeleton-Based Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2021, pp. 1915–1925, doi: 10.1109/TCSVT.2020.3015051.

[27] Shi, L.—Zhang, Y.—Cheng, J.—Lu, H.: Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12018–12027, doi: 10.1109/CVPR.2019.01230.

[28] Gao, X.—Hu, W.—Tang, J.—Liu, J.—Guo, Z.: Optimized Skeleton-Based Action Recognition via Sparsified Graph Regression. Proceedings of the 27th ACM International Conference on Multimedia (MM '19), 2019, pp. 601–610, doi: 10.1145/3343031.3351170.

[29] Yang, H.—Gu, Y.—Zhu, J.—Hu, K.—Zhang, X.: PGCN-TCA: Pseudo Graph Convolutional Network with Temporal and Channel-Wise Attention for Skeleton-Based Action Recognition. IEEE Access, Vol. 8, 2020, pp. 10040–10047, doi: 10.1109/ACCESS.2020.2964115.

[30] Hedegaard, L.—Heidari, N.—Iosifidis, A.: Continual Spatio-Temporal Graph Convolutional Networks. Pattern Recognition, Vol. 140, 2023, Art. No. 109528, doi: 10.1016/j.patcog.2023.109528.

[31] Liang, D.—Fan, G.—Lin, G.—Chen, W.—Pan, X.—Zhu, H.: Three-Stream Convolutional Neural Network with Multi-Task and Ensemble Learning for 3D Action Recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 934–940, doi: 10.1109/CVPRW.2019.00123.

Qichen Qin is currently studying for his Master's degree at the Taiyuan University of Technology. His research focuses on human action recognition and computer vision.

Xiaohong Han is Professor at the Taiyuan University of Technology and a Master's tutor. She visited the University of Texas at Dallas in the USA for a one-year scientific research exchange and cooperative research on the application of intelligent optimization algorithms. Her main research areas are big data mining, artificial intelligence, pattern recognition, image processing, etc.