IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 September 2023, accepted 6 November 2023, date of publication 9 November 2023,
date of current version 17 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3331687

RESEARCH ARTICLE

Contact Part Detection From 3D Human Motion Data Using Manually Labeled Contact Data and Deep Learning

CHANGGU KANG[1], MEEJIN KIM[2], KANGSOO KIM[3], (Member, IEEE), AND SUKWON LEE[2]
[1]School of Computer Science, Gyeongsang National University, Jinju-si 52725, South Korea
[2]Korea Electronics Technology Institute, Gyeonggi-do 13509, South Korea
[3]Department of Electrical and Software Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada

Corresponding author: Sukwon Lee (sukwonlee@keti.re.kr)

This work was supported in part by the Culture, Sports and Tourism Research and Development Program through the Korea Creative Content Agency funded by the Ministry of Culture, Sports and Tourism (Project Name: Development of an Intelligent Surveillance Platform and Building a Training Dataset for the Security of the Integrated Resort) (90%) under Project R2022020028; and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2022-03294.

ABSTRACT Research on the interaction between users and their environment has been conducted in various fields, including human activity recognition (HAR), human-scene interaction (HSI), computer graphics (CG), and virtual reality (VR). Typically, the interaction process commences with a human body part's movement and involves contact with a target object or the environment. The choice of the body part to make contact depends on the interaction's purpose and affordance, making contact a fundamental aspect of interaction. However, detecting the specific body parts in contact, especially in the context of 3D motion and complex environments, poses computational challenges. To address this challenge, this study proposes a method for contact detection using motion data. The motion data utilized in this study are limited to actions feasible in an office environment. Since contact states of different body parts are independent, the proposed method comprises two distinct models: a feature model generating common features for each body part and a part model recognizing the contact state of each body part. The feature model employs a bidirectional long-short term memory(Bi-LSTM) structure to capture the sequential nature of motion data, ensuring the incorporation of continuous data characteristics. In contrast, the part model employs separate weights optimized for each body part within the deep neural network. Experimental results demonstrate the proposed method's high accuracy, recall, and precision, with values of 0.99, 0.97, and 0.95, respectively.

INDEX TERMS Affordance, contact detection, human activity recognition, human-scene interaction, human motion.

I. INTRODUCTION

Human activities in our daily lives involve various physical interactions with the surrounding objects and environments, such as walking on the street, grabbing a door handle to open it, and sitting on a chair. Such interactions are initiated by the movements of the actor's body parts and are associated with physical contact between the body parts and target objects(or environments) [1]. Owing to the inevitability and necessity of physical contact with objects in human activities, detecting and analyzing contacts/touches in real-time during activities is crucial for designing context-appropriate, intelligent, and effective systems.

Despite their importance and potential benefits, identifying the locations of physical contacts and their moments in 3D environments remains a challenging problem that requires complex contextual data and advanced processing method such as wearable sensors and computer vision algorithms. Unsurprisingly, various methods for body-contact analysis have been proposed and evaluated in the fields of human activity recognition (HAR) and human-scene interaction (HSI) [2], [3], [4], [5], [6], [7]. These methods often

The associate editor coordinating the review of this manuscript and approving it for publication was Janmenjoy Nayak.

consider the physical affordances of a target object and 3D data to identify the interaction between the actor's body part and the object while utilizing or developing novel sensing and processing approaches, such as depth cameras, infrared (IR) cameras, inertial measurement units (IMUs), and light detection and ranging (LiDAR) sensors. 3D human behavioral motion data and environmental data have been collected and analyzed to understand the contact status of each part of the actor's body.

However, previous research has typically followed separate path for human motion capture [8], [9], [10], [11] and 3D scene reconstruction [12], [13], [14], [15] due to the high complexity of the motion data and the computational cost of processing algorithms. This separation results in a lack of mutual complementation between motion and environmental data, making comprehensive body contact analysis challenging. Furthermore, heterogeneous equipment and data involved in human activity analysis make the analysis even more difficult (or impossible). For instance, motion data may be represented as a body skeleton, excluding human surface/skin information, and IMU-sensor-based equipment cannot capture environmental information. In previous works, detecting contact with the environment or an object in 3D space required environmental data [16] or physical information [17]. Furthermore, as the complexity of the environment and the complexity of motion increase, detecting contact becomes even more challenging. To address these issues, there is a need for methods that can detect contact status using only motion data without the need for additional information.

To resolve this problem, we propose a method to detect contacted parts among segmented body parts (e.g., hands, feet, back, upper leg, and lower arm) using only motion data, without 3D environment data. The proposed method consists of two models: a feature model and a part model. The feature model creates a common feature to be used as the input of the part model. Given the continuous nature of time-series body movement data, the feature model is created based on the recursive neural network (RNN) structure, and the part model recognizes the contact state of each body part using the previous feature vector. All parts have the same structural model; however, each part has its own weight because the contact state of each part can be independent. The model for each part is based on a deep neural network (DNN) structure. Our main contribution is that the proposed method detects the contacted parts among human bodies by using only a motion data without the environment information and physical features such as weight and height of a human.

For the learning and experimentation of the proposed method, we collected basic body movements and actions that could occur in an office environment with context-appropriate objects such as telephones, water cups, cabinets, and chairs. Additionally, we manually mark the contact status in each frame of the collected motion data. Because of the dominant hand and walking motions in the office environment, most actions are associated with contact with the hands and feet; however, in some other cases, such as sitting on a chair, a complex contact state can be involved depending on the chair form and location.

The remainder of this paper is organized as follows. Section II describes research trends in interaction analysis and contact detection. Section III describes the data collection, feature configuration, and learning machine construction. Section IV describes the proposed method in detail. Section V evaluates its performance and explains the experimental results. Finally, Section VI concludes the paper with a discussion of its limitations and future studies.

II. RELATED WORK

A. CONTACT DETECTION BASED ON BODY MOTION DATA

Various studies have been conducted regarding contact. Foot-contact information increases the accuracy of motion capture by providing physical constraints. Ma et al. [17] conducted a study on foot contact detection methods. The proposed detection method includes both motion data and the actor's weight information because motion data alone are insufficient. Lee and Lee [16] suggested an iterative method for detecting body parts that contact with a complex environment. This method can deal with a highly dynamic motion by considering the physical properties, but the 3D environment information must be given. Narasimhaswamy et al. [18] proposed a method for detecting hand contact states in complex environments. The proposed method accurately identifies the positions of the contacting hands from input images using masked region-based convolutional neural network(R-CNN) [22] and an attentional pooling module [23], even in highly complex situations with multiple contacted hands. However, the proposed method operates on static images and lacks continuous contact information and the ability to distinguish between the left and right hands. Kang and Lee [24] allocated contactable candidate points on a character's surface and conducted research to generate appropriate poses to maintain stable postures in a given environment. Although it can generate various poses based on the user's input position and orientation, it differs from the proposed research in that the contact points are found in the environment. Several studies have been conducted on foot-contact detection. Kim and Lee [19] used physical sensors, whereas Ma et al. [17] utilized dynamic information and required both motion data and actor weight information. Most of these studies required additional information or devices for contact detection. Recent studies have explored finding environmental contacts using 3D mesh pose data. Hassan et al. [20] aimed to probabilistically visualize mesh data, showing the contact for each body part given 3D poses. This study shares a goal similar to ours of finding contacts from 3D data. They represented specific contact regions at the mesh level compared to our goal. However, they used static data as input data, whereas we found contacted parts from motion data, which is a strength of our study. Huang et al. [21] proposed a method for

TABLE 1. Comparison of studies on contact detection for human interaction.

Study	Main Focus	Data Input	Key Features/Strengths
Ma et al. [17]	Foot contact detection	Motion data, actor's weight	Considers physical constraints, dynamic motion
Lee and Lee [16]	Detecting body parts	Highly dynamic motion	Requires 3D environment information
Narasimhaswamy et al. [18]	Hand contact detection	Input images	Accurate identification, handles complex situations
Kim and Lee [19]	Foot contact detection	Physical sensors	Requires additional devices/information
Hassan et al. [20]	Mesh contact visualization	3D mesh pose data	Represents specific contact regions at mesh level
Huang et al. [21]	3D contact estimation	Single 2D image	Promising approach for contact information
Our study	Contact-labeled motion	Human motion	Use only human motion data without 3D environment and physical information

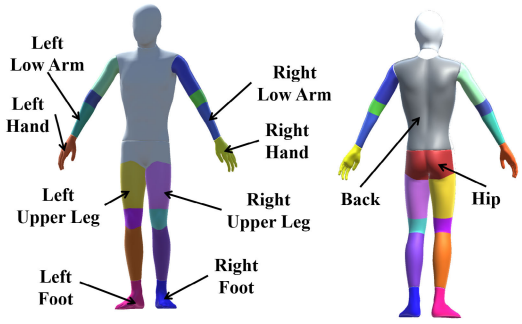

FIGURE 1. Segmented parts of the human body and contact colors.

estimating the 3D contact information on the human body from a single 2D image. Their method estimated the contact information on the vertex level without 3D reconstruction, making it a promising approach. However, contrary to our objective, these studies targeted static poses. Table 1 shows the comparison of studies on contact detection for human interaction.

B. CONTACT DETECTION WITH 3D ENVIRONMENTAL INFORMATION

To analyze interactive actions within the environment, both user poses and environmental information are required. However, acquiring such information in the environment is challenging owing to occlusion from interacting objects, which prevents complete information retrieval. To address this problem, Hassan et al. [25] proposed a method for estimating suitable poses using 3D geometric information from incompletely captured human pose data in an interactive environment. This method can accurately detect contact information between adjusted poses and the environment; however, it relies only on static pose information and has the limitation of requiring accompanying environmental data. Research related to the spatial representation in areas where interactions occur has also been conducted. Savva et al. [26] represented probabilistic interaction information in the environment using action maps reconstructed by 3D depth maps and proposed a method for finding feasible spaces for given actions. The most crucial information in the interaction is contact information, and representation through depth maps does not sufficiently reflect this contact information.

TABLE 2. Types and details of the collected behavioral data.

Motion #	Subject	Motion Details
M0	Moving	Moving forward, moving after a 180° rotation to the left, moving after a 90° rotation to the left, moving after a 180° rotation to the right, moving after a 90° rotation to the right.
M1	Sitting	Sit and stand by putting your hands on your desk
M2		Sit and stand up by placing your hands on your knees
M3		Sit down on the back of the arm and back on the back of the armrest
M4	Open the door and go in	After moving, pull the door with your right hand and open and enter
M5	Open and close the cabinet	Open the cabinet door with both hands and find a simple document and close the cabinet door with both hands
M6	Drinking water	After moving forward, drink from the cup on the table and place it in its original position.
M7	Answering the phone	Moving forward, picking up the handset on the table, having a conversation, and then placing the handset back in its original position.

To address this issue, Gupta et al. [27] proposed a method for clearly representing person-scene contact relationships in 3D scene data. Although this has the advantage of representing the relationships between actions as a graph, it differs from this study in that it uses 2D images instead of 3D data.

III. DATA COLLECTION

The equipment used to collect motion data can be broadly categorized into vision-based devices (such as optical and depth cameras) and devices that use IMU sensors. In our study, we utilize Perception Neuron, an IMU sensor-based device capable of capturing human motions effectively, even in the presence of occlusion caused by object movement [28]. Table 2 lists the types of motion data collected and their detailed descriptions. The walking motion, as the most common basic motion performed by humans, was captured in seven detailed actions, including straight walking and 90° turns (left and right). The "Sitting" motion involves various changes in contact states and encompasses a significant amount of contact occurrences depending on the type and

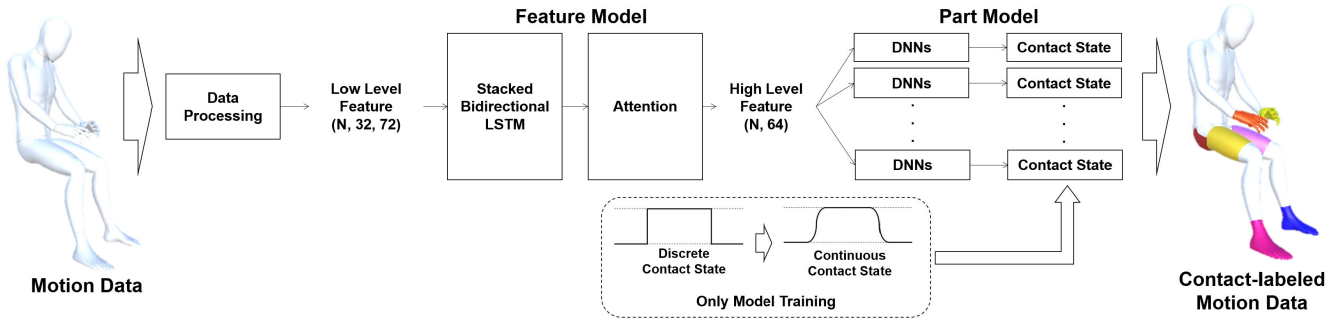

FIGURE 2. An overview of our method.

method of chairs. We used three types of chairs: desk chair, armchair, and stool. The detailed action of the armchair involves using the backrest and armrests while sitting, whereas the desk chair involves sitting using the knees. Finally, for the stool, the action involved pulling the desk with the hands and sitting down. The "Opening and Entering a Door" motion involves a hinged door, where the door handle is turned to open and enter. Cabinet-related motions included opening the cabinet with both hands, finding documents inside, and closing the cabinet again with both hands. The motions of "Drinking Water" and "Answering a Phone" involve moving to a target location and drinking water from a cup or answering a phone call. After completing the motion, the subject returns to the initial position. We collected 22 motion data for the same subject by repeating motion capture 22 times.

To facilitate training, both collected motion data and contact state data are required. To create contact-state data, the human body was divided into multiple parts, and 11 specific parts were chosen for contact-state recognition. The parts used for recognizing the contact states include the left (right) hand, left (right) foot, hips, back, upper left (right) leg, and lower left (right) arm. Figure 1 shows the segmented parts of the human body and the parts used for recognizing contact states. The parts for contact-state recognition are chosen arbitrarily based on their significance during office-related movements. The upper and lower legs are excluded because they are not involved in actual contact during office movements. Additionally, although the spine consists of three distinct segments (spine 1, spine 2, and spine 3) in the motion data, it is treated as a single entity for recognizing the contact states. The contact state for each frame in the motion data is manually labeled as 0 (non-contact) or 1 (contact).

IV. PROPOSED METHOD

Figure 2 shows an overview of the proposed method. The purpose of our method is to generate the contact-labeled motion data from the original motion data. The input motion data are represented in the form of a .bvh file that captures the angles between the bones forming the skeleton at specific time points. Through the data processing by using the motion data, a feature vector is created to encompass spatial-temporal

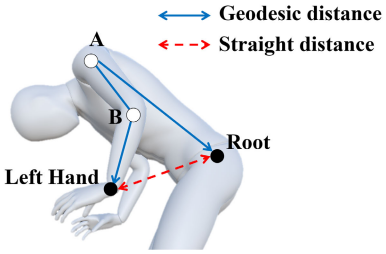

FIGURE 3. Straight distance and geodesic distance for RDR.

features such as joint distances, distance ratios, and angles between non-neighboring but significant bones. This feature vector is referred to as a low-level feature (LLF). Each LLF represents the features of every frame derived from the raw motion data. Subsequently, a two-dimensional (2D) vector is formed by incorporating continuous data from a specific sequence, which is then transformed into a high-level feature (HLF) through a feature model that determines the contact state in each part of the model. The feature model has a structure consisting of a stacked Bi-LSTM and an attention layer. While the feature model processes the data, the 2D LLF is converted into a 1D HLF and the 1D HLF is used as common features to determine the contact state of each part. Because we use deep a learning model, we need the contact data of each motion frame for the model training, which generally takes the form of discrete binaries. The data is transformed into a continuous form to minimize loss during the training process. Finally, the part model employs HLF to recognize the contact state of each part, with independent part model weights assigned to each part during the training process. The contact-labeled data was generated using the trained models.

A. DATA PROCESSING

The data consists of motion and contact data. The motion data represents the angles of the joints in the skeleton, and we create the LLF through data processing process. The LLF includes the following four values.

- Absolute distance (AD): The distance of each segment from the root.

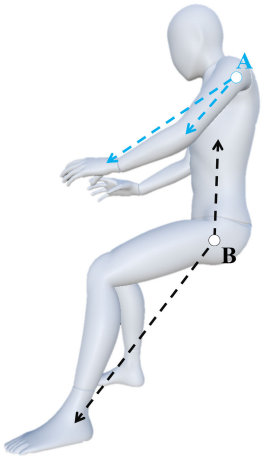

FIGURE 4. An example of PA: A represents the angle between the shoulder-to-wrist and shoulder-to-elbow vectors, whereas B represents the angle between the hip-to-spine and hip-to-ankle vectors.

- Relative distance rate (RDR): The ratio of the geodesic distance to the straight-line distance of each part from the root in (1).

$$RDR = \frac{Distance_{straight}}{Distance_{geodesic}} \tag{1}$$

- Angular velocity (AV): Angular velocity for each part
- Pair angle (PA): Angle of a pair of parts

Figure 3 shows an example of geodesic distance measurement. If we measure the geodesic distance from the root to the left hand, we calculate the distance along the skeletal structure, following the path: Root -> A -> B -> Left Hand.

Except for the pair angle (PA), the remaining values are obtained from the other 12 parts, excluding the head, as shown in Figure 1. Variations in the head during office movements are not specifically relevant, except for eye movements; therefore, they are excluded. AD and RDR each have a size of 12 $\times$1, while AV has a size of 36 $\times$1, as it calculates the angular velocities for each of the three axes. The PA includes the angles between the wrist and forearm, forearm and upper arm, knee and lower leg, and ankle and foot, as shown in Figure 4, resulting in a size of 12 $\times$1, incorporating the shoulder-to-hand, hip-to-thigh, and knee-to-ankle angles.

By concatenating these four values, the LLF has a vector of size 72 $\times$1 in (2).

$$LLF = concat(f_{AD},\ f_{RDR},\ f_{AV},\ f_{PA}) \tag{2}$$

During data processing, binary contact states represented by 0s and 1s are transformed into continuous states. In general, the contact states are represented as 0s and 1s. However, the values output by algorithms usually appear as decimal numbers between 0 and 1, and are rounded to achieve binarization. Specifically, values equal to or greater than 0.5, are considered as contact, while values less than 0.5 are considered as non-contact (release). In the sequence models

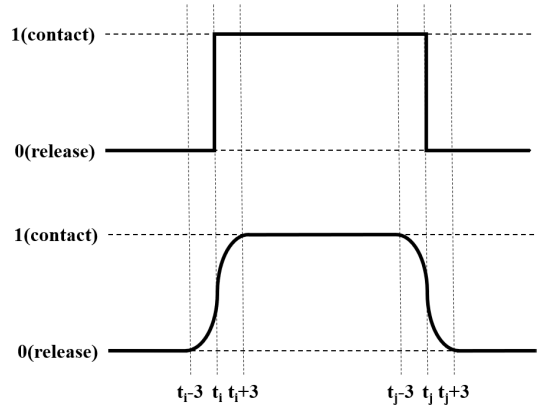

FIGURE 5. General contact graph and the contact graph applied for sigmoid function.

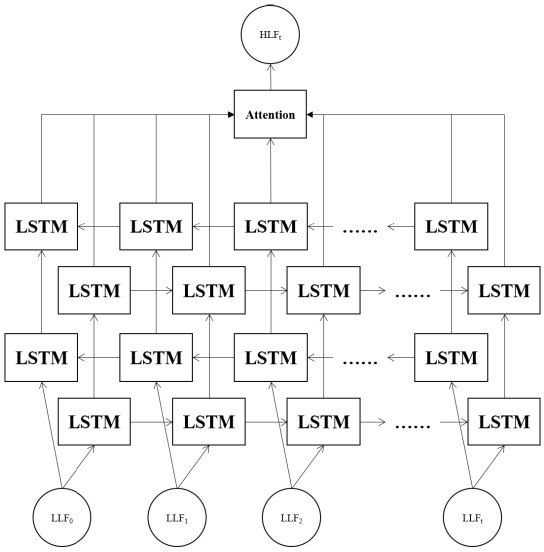

FIGURE 6. Feature model.

used for continuous data, rapid changes in the estimated output values create unstable contact-state graphs. To address this issue, we apply the sigmoid function to six frames, consisting of three frames before and after the point of contact change, based on the reference point of the contact transition. This process transforms the contact states into continuous contact state. In the contact state graph, when the contact state transitions from (0, 1) or (1, 0), (3) (left) and (3) (right) are used to create continuous contact state graphs. For the case of (1, 0), which represents the transition from 1 to 0, an inverted sigmoid formula is applied.

Figure 5 shows the discrete contact graph, which is represented by 0 and 1, and the continuous contact graph obtained by applying the sigmoid function. The transition of the contact state occurs at time points t_i and t_j, and a sigmoid function is applied to the $\pm$3 frames around each time point to create a continuous contact state.

$$\frac{1}{1+e^{i}},\ \frac{1}{1+e^{-j}} \tag{3}$$

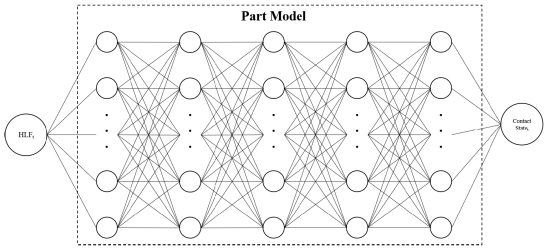

FIGURE 7. Part model to recognize the contact state for each part.

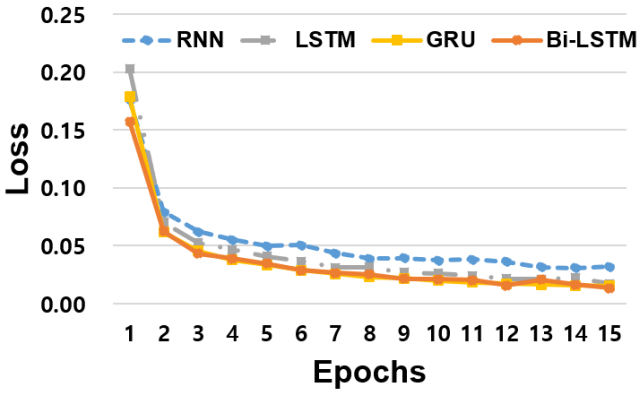

FIGURE 8. The graphs of training loss for RNN, LSTM, GRU, and Bi-LSTM.

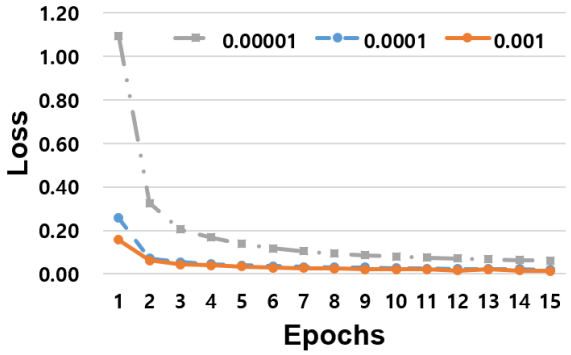

FIGURE 9. The graphs of training loss according to learning rate.

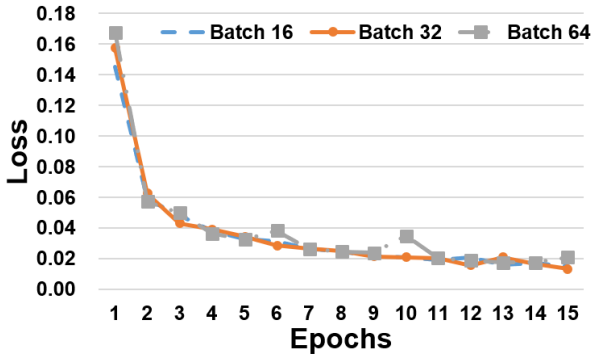

FIGURE 10. The graphs of training loss according to the size of batch.

B. MODEL DESIGN

The proposed method is divided into two components: a feature model that generates the HLF and a part model that recognizes the contact state of each part. The feature model has a recurrent neural network structure because the input LLF has a continuous value variation. In this case, Bi-LSTM structure [29] is employed, where both forward and backward learning of the motion data are utilized to incorporate the results into the HLF generation. To address the reduction in the size of the feature vector during HLF generation and mitigate information loss and the vanishing gradient problem, we add an attention layer [30]. The attention layer employs dot attention and calculates attention scores. The concatenation of the forward and backward hidden state values of the Bi-LSTM is multiplied by the input value of the layer through the dot product, and the SoftMax function is applied to determine the attention weights, ensuring that their sum is equal to 1. Finally, element-wise multiplication of the input value and attention weights is subjected to a reduced sum, resulting in an output size of 64×1. Figure 6 shows the structure of the feature model. The Bi-LSTM consists of four layers with an input size of 32×72 and an output size of 64×1.

The part model is designed to recognize the contact state of each part independently as each model has unique weights during a training process. And The model utilizes the HLF generated from the feature model to recognize the contact state for each part. It has a fully-connected layer (FCL) structure to ensure that suitable weights are computed for each feature vector element. The Part Model is composed of five layers, and each layer has an output size of 8×1. The final layer applies the sigmoid activation function, resulting in a single output value ranging from 0 to 1. The loss function is the mean squared error, and the Adam optimizer [31] is employed for optimization. Figure 7 shows the structure of the part model.

Figure 8 shows the graphs of training loss for RNN, LSTM, gated recurrent unit(GRU), and Bi-LSTM. Most graphs have a similar pattern graph. But the Bi-LSTM has the least loss comparing with others. And we conducted an experiment to find the optimal learning rate and batch size. Figure 9 and Figure 10 respectively show the results according to learning rate and batch size. The proposed model has the lowest loss when the learning rate is set to 0.001 and the batch size to 32. Figure 11 shows the experimental results for the number of Bi-LSTM layers being 3, 4, 5, and 6. We found that when the number of Bi-LSTMs was 4, the loss was minimum. Table 3 displays our model's details and settings for training. We have also configured the hyperparameters for training the proposed model, setting the sequence length of the training data to 32 and the number of epochs to 15.

V. EXPERIMENTS

We performed the experiments for contact detection using the trained model. Figure 12 shows the loss change when the binary and continuous contact states were applied separately. It can be observed that the loss is consistently lower for the continuous contact states at all intervals. Figure 13 shows the

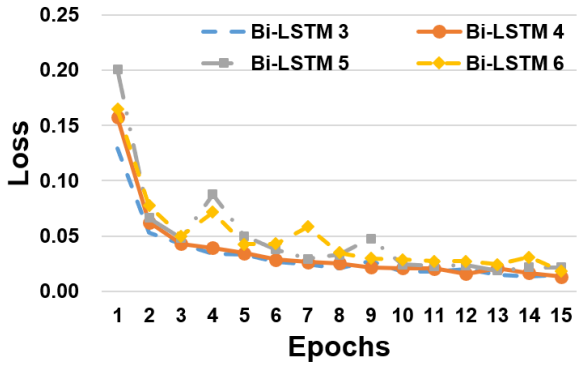

FIGURE 11. The graphs of training loss according to the number of Bi-LSTM layers.

TABLE 3. Model details and settings.

Component	Model Details
Feature Model	- Architecture: Bi-LSTM [29] - Layers: Four Bi-LSTM layers - Input Size: 32×72 - Output Size: 64×1 - Attention Mechanism: Dot attention with SoftMax
Part Model	- Architecture: Fully-connected layers - Layers: Five layers - Output Size per Layer: 8×1 - Final Layer Activation: Sigmoid function - Output Range: 0 to 1
Training	- Loss Function: Mean Squared Error (MSE) - Optimization: Adam optimizer [31] - Learning Rate: 0.001 - Sequence Length: 32 - Batch Size: 32

TABLE 4. Configuration of the confusion matrix for experiments.

		Recognized State	
		C	N
Actual State	C	CC	CN
	N	NC	NN

changes in the loss of each part during the training process. In the last 15 epochs, the loss for each part was all < 0.01, which is a very low value.

To evaluate the performance of the model, we present the recognition results in a confusion matrix, as shown in Table 4. In the confusion matrix, "contact" and "non-contact" are denoted as "C" and "U" respectively. The numbers in the confusion matrix represent the number of frames in each state and recognized state of the results. Based on the values in the confusion matrix, we calculated the accuracy, recall, and precision using (4), (5), and (6), respectively.

$$accuracy = \frac{CC + NN}{CC + NC + CN + NN} \tag{4}$$

$$recall = \frac{CC}{CC + NN} \tag{5}$$

$$precision = \frac{CC}{CC + NC} \tag{6}$$

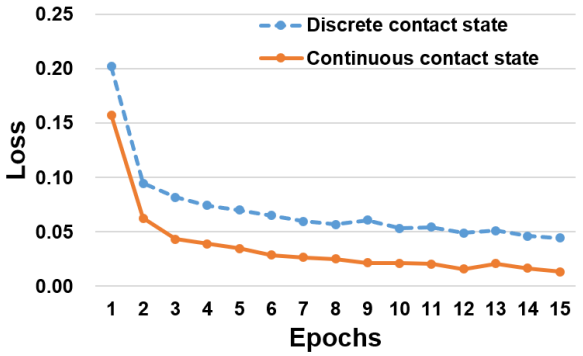

FIGURE 12. The graphs of training loss for the discrete contact state and continuous contact state.

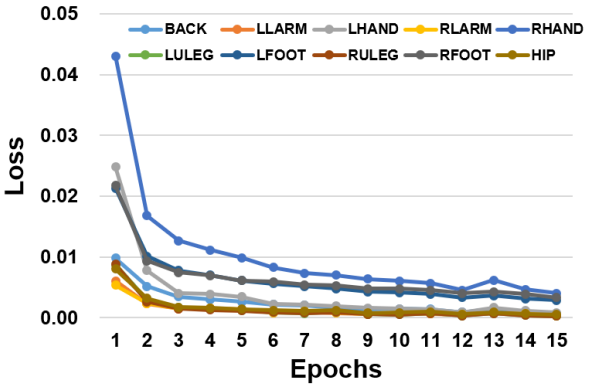

FIGURE 13. The graphs of training loss for each part.

Table 5 shows the confusion matrix, accuracy, recall, and precision for each part based on the recognized contact results. In the confusion matrix, all parts except for the foot start in the non-contact state, and the frequency of the non-contact state was relatively high because the actual interaction time was short. The total number of frames in all motion data used in the experiment was approximately 130K, with a contact proportion of approximately 24% (31,629 frames) and a non-contact proportion of 76% (98,761 frames). Table 6 shows the proportion of contact states for each part compared with the total contact states (32K). Because all data involved movements on the ground, the proportions of the Left Foot and the Right Foot accounted for over half of the total, and their proportions are similar. In addition, LeftLowArm and RightLowArm have lower proportions because contact occurs only when obtaining support from the armrest of a chair. The average accuracy of each part model was 0.99, and even for parts with a small difference between contact and non-contact, an accuracy of approximately 0.98 was obtained. The recall and precision showed a high performance (over 0.90 for most parts). However, the recall for the back part was lower at 0.85 compared to the other metrics and parts. This is because in the actual sitting posture (training data), contact occurs when the backrest starts touching, but in the recognition system, contact is recognized when the backrest is fully touched. Determination

TABLE 5. Confusion matrix for each part.

Part		Back		Left Low Arm		Left Hand		Right Low Arm		Right Hand		Left Upper Leg		Left Foot		Right Upper Leg		Right Foot		Hip	
Recognized State		C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
Actual State	C	427	80	230	4	2213	66	231	3	3525	109	1203	15	10725	36	1203	15	10397	73	1065	9
	N	21	12511	24	12781	25	10735	23	12782	164	9241	40	11781	262	2016	37	11784	211	2358	49	11916
Accuracy		0.99		1.0		0.99		1.0		0.98		1.0		0.98		1.0		0.98		1.0	
Recall		0.84		0.98		0.97		0.99		0.97		0.99		1.0		0.99		0.99		0.99	
Precision		0.95		0.90		0.98		0.90		0.95		0.96		0.97		0.97		0.98		0.95	

TABLE 6. Ratio of contact states for each part compared to the total contact state.

Part	Back	Left Low Arm	Left Hand	Right Low Arm	Right Hand
%	1.6	0.73	7.2	0.74	11.5
Part	**Left Upper Leg**	**Left Foot**	**Right Upper Leg**	**Right Foot**	**Hip**
%	3.9	34.0	3.9	33.1	3.4

FIGURE 14. The results of applying contact recognition to motions and visualizing them: From the top to bottom, the contact recognition results for M0-M3 motions are shown.

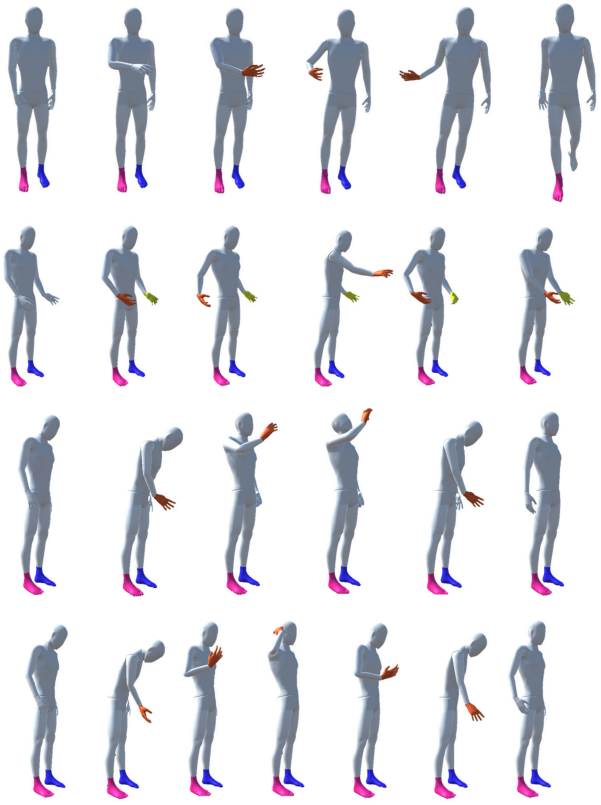

FIGURE 15. The results of applying contact recognition to motions and visualizing them: From the top to bottom, the contact recognition results for M4-M7 motions are shown.

of the contact status based solely on motion data when the person is sitting with their backs upright, rather than leaning against the backrest of the chair, is challenging.

Figure 14 shows a visualization of the contact parts extracted from the motion(M0-M3). When contact occurs in each part, the colors shown in Figure 1 appear. M0 is a walking motion, so it can be observed that contact alternates between the left and the right feet. M1 is a motion of sitting down and standing up with the hands on the desk. While a person in the motion interacts with a chair and a desk, contact occurs when the hands touch the desk while sitting and standing up from a chair, followed by the UpperLeg and Hips. And both feet keep the contact on the floor. M2 is a sitting motion, and a person in M2 uses a backrest and his own knees to sit. In cases where there is a backrest on the chair, but the motion does not fully utilize the backrest, it can be ambiguous to confirm contact because the person's back and thighs are nearly in a vertical posture, and there is no pressure applied to the backrest. Therefore, to ensure

TABLE 7. Confusion matrix for each motion.

Motion #		Back		Left Low Arm		Left Hand		Right Low Arm		Right Hand		Left Upper Leg		Left Foot		Right Upper Leg		Right Foot		Hip	
M0		C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
Actual	C	0	0	0	0	0	0	0	0	0	0	0	0	2308	10	0	0	2223	33	0	0
State	N	0	3428	0	3428	0	3428	0	3428	0	3428	0	3428	98	1012	0	3428	77	1095	0	3428
Accuracy		N/A		N/A		N/A		N/A		N/A		N/A		0.97		N/A		0.97		N/A	
Recall		N/A		N/A		N/A		N/A		N/A		N/A		1.0		N/A		0.99		N/A	
Precision		N/A		N/A		N/A		N/A		N/A		N/A		0.96		N/A		0.97		N/A	
M1		C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
Actual	C	0	0	0	0	435	15	0	0	433	15	259	0	1299	2	259	0	1252	6	240	4
State	N	0	1562	0	1562	3	1109	0	1562	5	1109	24	1279	39	222	23	1280	31	273	18	1300
Accuracy		N/A		N/A		0.99		N/A		0.99		0.98		0.97		0.99		0.98		0.99	
Recall		N/A		N/A		0.97		N/A		0.97		1.0		1.0		1.0		1.0		0.98	
Precision		N/A		N/A		0.99		N/A		0.99		0.92		0.97		0.92		0.98		0.93	
M2		C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
Actual	C	193	51	0	0	504	19	0	0	494	29	496	2	1370	14	495	3	1323	21	398	3
State	N	0	1346	0	1590	10	1057	0	1590	6	1061	13	1079	13	193	13	1079	14	232	21	1168
Accuracy		0.97		N/A		0.98		N/A		0.98		0.99		0.98		0.99		0.98		0.98	
Recall		0.79		N/A		0.96		N/A		0.94		1.0		0.99		0.99		0.98		0.99	
Precision		1.0		N/A		0.98		N/A		0.98		0.97		0.99		0.97		0.99		0.95	
M3		C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
Actual	C	234	29	230	4	558	6	231	3	709	6	448	13	1480	8	449	12	1451	10	427	2
State	N	21	1463	24	1489	0	1183	23	1490	26	1006	3	1283	42	217	1	1285	18	268	10	1308
Accuracy		0.97		0.98		1.0		0.99		0.98		0.99		0.97		0.99		0.98		0.99	
Recall		0.89		0.98		0.99		0.99		0.99		0.97		0.99		0.97		0.99		1.0	
Precision		0.92		0.91		1.0		0.91		0.96		0.99		0.97		1.0		0.99		0.98	
M4		C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
Actual	C	0	0	0	0	0	0	0	0	168	22	0	0	699	1	0	0	648	1	0	0
State	N	0	866	0	866	0	866	0	866	7	669	0	866	34	132	0	866	34	183	0	866
Accuracy		N/A		N/A		N/A		N/A		0.97		N/A		0.96		N/A		0.96		N/A	
Recall		N/A		N/A		N/A		N/A		0.88		N/A		1.0		N/A		1.0		N/A	
Precision		N/A		N/A		N/A		N/A		0.96		N/A		0.95		N/A		0.95		N/A	
M5		C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
Actual	C	0	0	0	0	716	26	0	0	385	37	0	0	1256	1	0	0	1246	0	0	0
State	N	0	1430	0	1430	12	676	0	1430	43	965	0	1430	26	147	0	1430	21	163	0	1430
Accuracy		N/A		N/A		0.97		N/A		0.94		N/A		0.98		N/A		0.99		N/A	
Recall		N/A		N/A		0.96		N/A		0.91		N/A		1.0		N/A		1.0		N/A	
Precision		N/A		N/A		0.98		N/A		0.89		N/A		0.97		N/A		0.98		N/A	
M6		C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
Actual	C	0	0	0	0	0	0	0	0	546	0	0	0	1030	0	0	0	997	0	0	0
State	N	0	1082	0	1082	0	1082	0	1082	20	516	0	1082	4	48	0	1082	10	75	0	1082
Accuracy		N/A		N/A		N/A		N/A		0.98		N/A		1.0		N/A		0.99		N/A	
Recall		N/A		N/A		N/A		N/A		1.0		N/A		1.0		N/A		1.0		N/A	
Precision		N/A		N/A		N/A		N/A		0.96		N/A		0.99		N/A		0.99		N/A	
M7		C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
Actual	C	0	0	0	0	0	0	0	0	790	0	0	0	1283	0	0	0	1257	2	0	0
State	N	0	1334	0	1334	0	1334	0	1334	57	487	0	1334	6	45	0	1334	6	69	0	1334
Accuracy		N/A		N/A		N/A		N/A		0.96		N/A		1.0		N/A		0.99		N/A	
Recall		N/A		N/A		N/A		N/A		1.0		N/A		1.0		N/A		1.0		N/A	
Precision		N/A		N/A		N/A		N/A		0.93		N/A		0.99		N/A		1.0		N/A	

clarity in contact status, we encourage the complete use of the backrest. As a person in M3 interacts with a chair with a backrest and armrests, he uses them to sit comfortably. M3 is the motion of leaning against the backrest while sitting, and it can be observed that the chair armrest, seat, and backrest are sequentially in contact.

Figure 15 shows a visualization of the contact parts for the motion(M4-M7) using only hands. M4 is a motion that involves entering an office by opening a door. It can be observed that the right hand makes contact with the doorknob while the person opens the door. M5 is a motion that involves opening a cabinet with both hands and touching a book. After opening the cabinet, the right hand of a person reaches into the cabinet to touch a file. During this action, the left hand maintains contact with the cabinet door until just before closing it. For M5, the result on second line in Figure 15 shows that the contact occurs in order of interaction with a cabinet and a file. M6 is a motion involving the interaction with cups to drink water. The third motion in Figure 15 shows the contact parts during the drinking action. The contact of the right hand begins when it grasps the cup and continues until just before placing the cup back after drinking. M7 is a motion

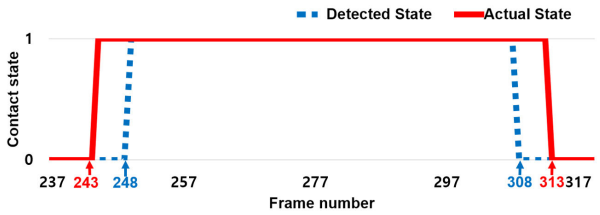

FIGURE 16. Contact state graph for the back part of M2.

involving answering a phone. The fourth motion shows the contact parts for M7. The contact of the right hand starts from grasping a receiver. And the contact continues during talking on the phone. As soon as the person places the receiver back, the contact of a right hand ends.

Table 7 shows the confusion matrices for each type of motion. Except for BACK in M2 and M3, all values showed high results at 0.9 or higher. There were no cases in which non-contact state-only situations were incorrectly recognized as contact states in all motions and parts. Although most parts showed high recognition rate, there were differences in the recognition rate depending on the sitting behavior type of the back. This is because M1, M2, and M3 all performed the sitting; however, M1 performed the motion without back contact, whereas M2 and M3 performed the motion with back contact. In the case of M2, the start of the back contact was faster than that in the actual data owing to the influence of M1, while the end of the contact was faster in the actual data. Figure 16 shows the recognized contact state of the back part for the M2 motion and the actual contact state. In the actual data, it started at frame 243 and ended at 313, but in the recognized contact state, it started at a shorter frame, 248, and ended at 308.

VI. CONCLUSION

In this paper, we propose a method for detecting the contacted parts from 3D human motion data. The proposed method consists of two models: a feature model and a part model. The feature model utilizes a Bi-LSTM structure to process the sequential features of the motion data. The part model employs a DNN with optimized weights for each body part. Through experiments, we evaluated the accuracy, recall, and precision of the proposed method, resulting in values of 0.99, 0.97, and 0.95, respectively.

The limitation of our work is the time-consuming process of creating labeled motion data with contact tags for training. We collected the training data by manually tagging each frame, which is an inefficient method, necessitating alternative approaches. Additionally, because we rely on joint angles to recognize contact states, we cannot precisely determine the contact range for each body part. As a future research direction, we aim to expand the recognition of contact states to a wider range of behaviors beyond office interactions and further investigate methods for user behavior recognition using continuous contact states. Contact is an essential element of interaction, and we anticipate that contact information can be applied to various topics within the fields of HAR and HSI.

REFERENCES

[1] H. Yi, C. P. Huang, D. Tzionas, M. Kocabas, M. Hassan, S. Tang, J. Thies, and M. J. Black, "Human-aware object placement for visual environment reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3949–3960.

[2] H. Kjellström, J. Romero, and D. Kragić, "Visual object-action recognition: Inferring object affordances from human demonstration," *Comput. Vis. Image Understand.*, vol. 115, no. 1, pp. 81–90, Jan. 2011.

[3] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 14–29, Jan. 2016.

[4] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, Jul. 2013.

[5] Z. Liu, S. Wu, S. Jin, S. Ji, Q. Liu, S. Lu, and L. Cheng, "Investigating pose representations and motion contexts modeling for 3D motion prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 681–697, Jan. 2023.

[6] X. Li, S. Liu, K. Kim, X. Wang, M.-H. Yang, and J. Kautz, "Putting humans in a scene: Learning affordance in 3D indoor environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12360–12368.

[7] L. Zhang, W. Du, S. Zhou, J. Wang, and J. Shi, "Inpaint2Learn: A self-supervised framework for affordance learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3778–3787.

[8] N. Hesse et al., "Learning an infant body model from RGB-D data for accurate full body motion analysis," in *Proc. 21st Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Granada, Spain: Springer, Sep. 2018, pp. 792–800.

[9] M. Munaro, C. Lewis, D. Chambers, P. Hvass, and E. Menegatti, "RGB-D human detection and tracking for industrial environments," in *Proc. 13th Int. Conf. Intell. Auton. Syst. (IAS)*. Padova, Italy: Springer, Jul. 2014, pp. 1655–1668.

[10] J. Han, E. J. Pauwels, P. M. de Zeeuw, and P. H. N. de With, "Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment," *IEEE Trans. Consum. Electron.*, vol. 58, no. 2, pp. 255–263, May 2012.

[11] N. Hesse, S. Pujades, M. J. Black, M. Arens, U. G. Hofmann, and A. S. Schroeder, "Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2540–2551, Oct. 2020.

[12] K. Wang, G. Zhang, and H. Bao, "Robust 3D reconstruction with an RGB-D camera," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4893–4906, Nov. 2014.

[13] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss, "ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 7855–7862.

[14] H. Xu, J. Hou, L. Yu, and S. Fei, "3D reconstruction system for collaborative scanning based on multiple RGB-D cameras," *Pattern Recognit. Lett.*, vol. 128, pp. 505–512, Dec. 2019.

[15] W. Dai, Y. Zhang, P. Li, Z. Fang, and S. Scherer, "RGB-D SLAM in dynamic environments using point correlations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 373–389, Jan. 2022.

[16] S. Lee and S.-H. Lee, "Projective motion correction with contact optimization," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 4, pp. 1746–1759, Apr. 2019.

[17] H. Ma, W. Yan, Z. Yang, and H. Liu, "Real-time foot-ground contact detection for inertial motion capture based on an adaptive weighted naive Bayes model," *IEEE Access*, vol. 7, pp. 130312–130326, 2019.

[18] S. Narasimhaswamy, T. Nguyen, and M. H. Nguyen, "Detecting hands and recognizing physical contact in the wild," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7841–7851.

[19] M. Kim and D. Lee, "Development of an IMU-based foot-ground contact detection (FGCD) algorithm," *Ergonomics*, vol. 60, no. 3, pp. 384–403, Mar. 2017.

[20] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black, "Populating 3D scenes by learning human-scene interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14703–14713.

[21] C. P. Huang, H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovsky, D. Scharstein, and M. J. Black, "Capturing and inferring dense full-body human-scene contact," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13264–13275.
[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Georgia, Oct. 2017, pp. 2980–2988.
[23] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.
[24] C. Kang and S.-H. Lee, "Environment-adaptive contact poses for virtual characters," *Comput. Graph. Forum*, vol. 33, no. 7, pp. 1–10, Oct. 2014.
[25] M. Hassan, V. Choutas, D. Tzionas, and M. Black, "Resolving 3D human pose ambiguities with 3D scene constraints," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2282–2292.
[26] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner, "SceneGrok: Inferring action maps in 3D environments," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 1–10, Nov. 2014.
[27] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2012–2019.
[28] R. Sers, S. Forrester, E. Moss, S. Ward, J. Ma, and M. Zecca, "Validity of the perception neuron inertial motion capture system for upper body motion analysis," *Measurement*, vol. 149, Jan. 2020, Art. no. 107024.
[29] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[30] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.
[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

CHANGGU KANG received the B.S. degree in computer engineering from Pusan National University, in 2008, and the M.S. and Ph.D. degrees in computer science from the Gwangju Institute of Science and Technology (GIST), in 2010 and 2017, respectively. He is currently an Associate Professor with the School of Computer Science, Gyeongsang National University (GNU). Before joining GNU, he was a Postdoctoral Researcher with the Graduate School of Culture Technology, KAIST, South Korea. His research interests include character animation, machine learning, motion planning, VR/AR, and interactive media.

MEEJIN KIM received the Bachelor of Science degree in engineering and in computer science from Yonsei University, in 2016, and the Master of Science degree in engineering from the Korea Advanced Institute of Science and Technology (KAIST) (Advisor: Dr. Sung-Hee Lee), in 2019. She is currently a Researcher with the Korea Electronics Technology Institute (KETI). She studied virtual telepresence character animation with the Graduate School of Culture Technology. She has been with the VR/AR Research Center, since 2020, researching virtual character motion and interaction with the environment.

KANGSOO KIM (Member, IEEE) received the B.S. and M.S. degrees in electronics and computer engineering from Hanyang University and the Ph.D. degree in computer science from the University of Central Florida, in 2018. He is currently an Assistant Professor with the Department of Electrical and Software Engineering, University of Calgary. He was a Postdoctoral Researcher with the University of Central Florida with an appointment with the Institute for Simulation and Training and the College of Nursing (2019–2020) and the University of Delaware, in 2021. His research interests include pervasive context-aware eXtended reality (XR) systems, intelligent social interactions, and perception and cognition in XR.

SUKWON LEE received the M.Eng. degree in character animation from the School of Information and Mechatronics, Gwangju Institute of Science and Technology (GIST), South Korea, in 2013, and the Ph.D. degree in character animation from the Korean Advanced Institute of Science and Technology (KAIST) (Advisor: Sung-Hee Lee), in 2019. Since then, he has been with the Korean Electronic Technology Institute (KETI) as a Senior Researcher. He is involved with the Communications and Media Research and Development Division, VR/AR Research Center. His research interests include the generation of animated characters and human pose estimation for VR/AR applications.