IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 31, 2021, accepted April 25, 2021, date of publication May 4, 2021, date of current version May 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077487

Ensemble Three-Stream RGB-S Deep Neural Network for Human Behavior Recognition Under Intelligent Home Service Robot Environments

YEONG-HYEON BYEON[1], DOHYUNG KIM[2], JAEYEON LEE[2], AND KEUN-CHANG KWAK[1], (Member, IEEE)

[1]Department of Control and Instrumentation Engineering, Chosun University, Gwangju 61452, South Korea
[2]Intelligent Robotics Research Division, Electronics and Telecommunications Research Institute, Daejeon 34129, South Korea

Corresponding author: Keun-Chang Kwak (kwak@chosun.ac.kr)

This work was supported in part by the Information and Communication Technology (ICT) Research and Development Program of the Ministry of Science and ICT (MSIT)/Institute for Information and communication Technology Planning and evaluation (IITP) of Republic of Korea (Development of Human-Care Robot Technology for Aging Society) under Grant 2017-0-00162 (70%), and in part by the Basic Science Research Program through the National Research Foundation of Republic Korea (NRF) funded by the Ministry of Education under Grant 2017R1A6A1A03015496 (30%).

ABSTRACT This paper presents a method for recognizing behaviors in videos based on the ensemble RGB-S deep neural network, which combines RGB images and skeleton features from an action recognition database built in intelligent home service robot environments. The ensemble model is designed using the three-stream approach. The first stream classifies behaviors in videos using a convolutional neural network (CNN) based on a pre-trained ResNet101 model, which uses two-dimensional (2D) sequence images of actions as its input, and training a long short-term memory (LSTM) neural network with the sequence (RGB 2D-CNN + LSTM). The second stream directly manages the video and uses a three-dimensional (3D) CNN to include both temporal and spatial information. The 3D CNN is based on a pre-trained R3D-18 model (RGB 3D-CNN). The last stream uses the pose evolution image (PEI) method, which converts the skeleton sequence into a single-color image. The converted images are used as the input for the CNN (Skeleton PEI-2D-CNN). This approach not only reflects the spatial and temporal features of the behaviors in videos, but also includes all characteristics of the 2D sequence images, 3D videos, and skeleton sequences. Finally, a large-scale database for behavior recognition in videos, known as ETRI-Activity3D, is used in this study to verify the performance of the proposed deep neural network. A recognition performance of 93.2% is achieved in a cross-subject experiment, verifying the superiority of this method over models from previous studies.

INDEX TERMS Ensemble RGB-S deep neural network, ETRI-Activity3D database, human behavior recognition, transfer learning.

I. INTRODUCTION

Deep learning is an artificial intelligence (AI) technology that enables computers to think and learn like humans. By enabling robots to learn independently and combine the information they learned, deep learning offers new possibilities in the human–robot interaction (HRI) field. Beyond mimicking human expressions and behaviors, robots that identify solutions by practicing while repeatedly encountering mistakes have emerged because of deep learning technology. Robots for homes, industries, and disaster relief are being developed successively using AI technology and various sensors that can detect the position and gradient, as well as visual, auditory, and tactile senses. HRI is a multidisciplinary research field involving human–computer interactions, AI, robotics, natural language processing, and social science. Moreover, HRI is a key technology that can enable intelligent home service robots to interact naturally with users using a robot camera, microphone, and various sensors. The fact that these robots possess autonomous movement capacity, bi-directional properties of interaction, and various levels of control indicates that HRI possesses features different from those of conventional human–computer interactions. HRI includes various technologies, such as facial recognition, gesture recognition, behavior recognition, voice recognition, speaker recognition, sound localization, and sound

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

separation. Among these HRI technologies, behavior recognition in videos was emphasized in this study. Studies regarding behavior recognition have been performed previously, as follows.

Kim [1] proposed an HRI framework for recognizing a user's position, identification, and behaviors while a robot is interacting with the user. This framework, which includes several individual HRI components, can yield reliable data regarding the user and combines them effectively to respond to the demand of HRI service applications. Le [2] addressed the social HRI problem by suggesting a method for integrating a deep neural network with a kinematic robot system to render HRI behaviors more robust. Zhu [3] proposed a technology for natural HRIs to recognize hand gestures and daily activities based on the motion data and hierarchical hidden Markov model (HMM) in a smart assisted-living system for the elderly and disabled. Manzi [4] proposed a mobile robot capable of recognizing daily activities using depth cameras and wearable sensors. Manzi [4] suggested a method to conflate data of different characteristics using various sensing strategies.

Jang [5] proposed a four-stream adaptive convolutional neural network (CNN) that is robust to spatio–temporal changes such that daily activities can be recognized under robotic environments for elder care. Simonyan [6] proposed a behavior recognition method using two streams from images and an optical flow. One stream inputs individual frames, whereas the other stream uses several frames to compute the optical flow. Subsequently, they are input to the respective CNNs, and their scores are added to the final stage. Next, a two-dimensional (2D) convolution is performed, where 2D data are used as the input and a 2D result is generated; additionally, a three-dimensional (3D) convolution is performed in three directions. Hence, the 3D convolution uses 3D data as the input and generates a 3D result. Tran [7] proposed a 3D CNN based on 3D convolution operations for behavior recognition in videos. The CNN comprised eight convolution layers and two fully connected layers.

Wang [8] used the trajectory of body parts to classify actions performed. After extracting the trajectory from the videos, features extracted directly from the Fisher vector were combined with features learned through CNN-based deep learning in the final layer. Yang [9] proposed a multimodal combination comprising four models to classify videos. The four models comprised features of 3D convolution, 2D optical flows, 3D optical flows, and 2D convolution. The combination method used in the study was the boosting mechanism.

Zhang [10] proposed a dual-channel deep CNN that uses both a spatial channel network and a temporal channel network to extract static and dynamic features.

Wang [11] introduced pattern-based, model-based, and deep learning-based behavior recognition using channel state information. Gong [12] proposed a method to combine a pretrained deep learning model with a typical HMM for recognition in HRIs. Liu [13] suggested a deep fully connected model for group activity recognition. A spatio–temporal model based on CNN and long short-term memory (LSTM) networks was used to capture dynamic actions of individuals, and a fully connected conditional random field was used to learn the interactions between people.

Karpathy [14] suggested combining two streams in parallel for video classification. Two parallel encoders were constructed to be smaller and simplified using fewer parameters. One of the encoders managed low-resolution features, whereas the other processed high-resolution features. Subsequently, they were combined in the last stage of the fully connected layer. Ng [15] proposed two methods for classifying lengthy videos. The first method uses max pooling for the convolution features on the time axis, and the second method connects the convolution features using LSTM to process videos of varying lengths. The motion of objects in the video provides useful information regarding the action being performed, and the proposed method measures the object's motion using an optical flow.

Based on the literature review provided above, it is clear that previous studies of HRIs were performed using mobile, home service, and eldercare robots for behavior recognition. In addition, deep learning technology comprising multiple streams, in which skeleton, RGB, depth images are used as input, was primarily investigated as behavior recognition technology. Furthermore, techniques involving the CNN, LSTM, optical flow, and HMM have been applied.

Previous studies suggest that a large-scale behavior recognition database is to be developed and made publicly available. Moreover, investigations should be conducted using the database to develop deep learning-based behavior recognition technology with a high recognition rate such that intelligent home service robots can aid the elderly in performing daily tasks.

Therefore, in this study, a large-scale behavior recognition database was constructed, and an ensemble three-stream deep network was designed using the database to perform behavior recognition for HRIs under an intelligent home service robot environment. In this regard, the Electronics and Telecommunications Research Institute (ETRI) constructed the ETRI-Activity3D database, which is a 3D image dataset for recognizing the daily activities of elderly people and adults under robotic environments. These data were collected legally and securely with the approval of the IRB (Institutional Review Board) and after obtaining consent from the elderly and adults for the collection and use of their personal information. The proposed deep network for behavior recognition comprised a three-stream deep network based on 2D sequence images, 3D videos, and skeleton data. This approach considers both the 2D and 3D characteristics investigated in previous studies, and it retains the characteristics of skeleton changes.

This paper is organized as follows: Section 2 describes the combined model comprising the LSTM and 2D-CNN, a component of the three-stream RGB-S DNN, the 3D-CNN, and the skeleton-based pose evolution image (PEI)-2D-CNN. Section 3 describes the ensemble DNN and the late fusion

method. Section 4 provides the experimental results and consideration of the proposed ensemble DNN using the large-scale ETRI-Activity3D behavior recognition database. Finally, Section 5 presents the conclusion of the study.

II. THREE-STREAM DEEP NEURAL NETWORK

RGB videos that display behaviors are data of multiple RGB images, which are captured successively in a certain time interval. Because the captured images are shown successively, the objects in the video appear as if they are moving. The size of the data for RGB videos varies significantly based on the image resolution. The capacity of RGB videos is typically several times larger than those of other types of data, and they include various information, including those regarding the surrounding objects and circumstances. Meanwhile, skeleton data only contain information regarding the coordinates of the joints. Hence, the data size is small and it contains only information regarding the skeleton structure of the person. Although a person's skeletal movement provides essential information for behavior recognition, using only skeletal information is insufficient for distinguishing between similar behaviors. In such cases, the surrounding information may be required to recognize behaviors. The two types of data, RGB videos and skeleton data, possess different and unique characteristics. Therefore, designing an ensemble three-stream deep network that uses these two types of data as the input can generate a better synergy effect. This section describes the components of the ensemble three-stream deep neural networks: RGB 2D-CNN+LSTM, RGB 3D-CNN, and Skeleton PEI-CNN.

A. RGB 2D-CNN + LSTM

Because RGB videos have a 3D structure, in which 2D images are stacked in layers along the time axis, it is difficult to apply RGB videos to 2D CNNs. Therefore, 2D-CNNs are used to create a feature vector representing the spatial information of each frame of the 2D image. Subsequently, the feature vector enters LSTM as an input to extract and classify the temporal information features. LSTM is an efficient neural network for processing sequence data because it refers to the numerical values from the previous layers to compute the output. As a feature extractor, it can be applied to the 2D-CNN using various models, including newly designed and pre-trained models. Pre-trained models are based on studies by renowned scholars worldwide, and the performances of these models have already been verified. Pre-trained models are useful because they provide excellent performances and can reduce the effort required to define tens or hundreds of layers. Various pre-trained models exist, ranging from light and relatively shallow models to heavy and deep models. AlexNet [16], GoogLeNet [17], ResNet [18], and DenseNet [19] are examples of pre-trained models. GoogLeNet, which has demonstrated favorable performances for 2D-CNNs, was used in this study. It is a preconfigured model comprising nine inception modules. An inception module is a group of layers that perform

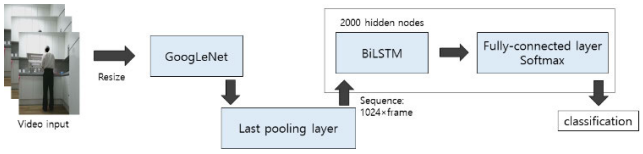


FIGURE 1. Diagram of RGB 2D-CNN + LSTM.

convolution operations of various sizes in parallel and combine the results to increase the neural network's depth effectively. The inception module reduces the number of feature maps by performing 1×1 convolution operations and considers minute features of the smallest units. Furthermore, it includes convolutions of various dimensions, such as 3×3 and 5×5. The contents of the previous layer are summarized through parallel pooling, and all parallel results are connected in the last step. Fig. 1 shows a diagram of the LSTM and 2D-CNN feature extractor that uses the RGB video sequence as the input.

The method for designing a deep neural network that classifies behavior recognition in videos is summarized as follows:

[Step1] Use a pre-trained CNN such as GoogLeNet to convert the videos into a sequence of feature vectors, and extract the features for each frame in the last layer pooling layer.

[Step2] Predict the video labels by training a bi-directional long short-term memory (BiLSTM) neural network with 2000 hidden units for the sequence.

[Step3] Design a neural network that classifies the videos directly by combining the layers of the two neural networks.

B. RGB 3D-CNN

In RGB videos, 2D images are stacked along the time axis, thereby forming a 3D structure. Therefore, it is difficult to apply RGB videos to 2D CNNs. In recent years, 3D-CNNs have outperformed 2D-CNNs that have been trained on large-scale video datasets. Because a 3D-CNN contains 3D filters, both temporal and spatial data are considered. Convolution operations and subsampling involve 3D filters, and other components involved are identical to those used in 2D-CNNs. Hence, favorable performance can be achieved using pre-trained models. The pre-trained models are C3D [7] and I3D [20] based on GoogLeNet, and R3D [21] based on ResNet. Deep layers cause some problems, where the gradient becomes excessively small or large and the performance is degraded as the layer becomes deeper. Hence, ResNet used skip connection, which reuses the input features of the previous layer. ResNet consists of five residual blocks, uses one of them as an input, and stacks the remaining with several overlaps [18]. In this study, behavior recognition was performed based on R3D-18. Furthermore, the codes used and the pre-trained models are publically available on GitHub [22]. Fig. 2 shows the simplified structure of the 3D-CNN with an RGB video input.

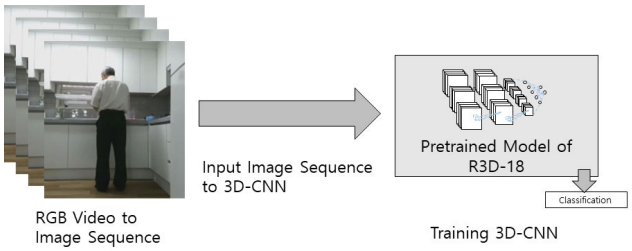


FIGURE 2. Architecture of RGB 3D-CNN.

C. SKELETON PEI-BASED CNN

Skeleton refers to human skeletal information extracted from sensor data. These data comprise the coordinates of joints, such as head, shoulder, hands, and feet, and they are defined for each frame to form a skeleton sequence. A skeleton is a data format that can effectively store the movements of a person of interest, and skeletons captured from a sensor are coordinates to form human skeleton. Similar to images, the skeleton of a certain moment does not contain all information regarding a person's action. Therefore, similar to videos, skeletons of several moments are arranged in a chronological sequence for behavior recognition. To effectively analyze this sequence data, both spatial and temporal data are important. Hence, conversion methods have been investigated to extract the appropriate information effectively. In this study, PEI, a method for converting a skeleton sequence into a single-color image, was used [23]. First, a joint is a central axis over which the body parts can be bent. Because ordinary people have a limited number of bendable joints, a human skeleton can be represented with minimal data. Kinect v1 represents a human skeleton using 20 joints. Meanwhile, Kinect v2, used in this study, uses 25 joints to represent a human skeleton. Using many joints, detailed skeletal changes of a person can be detected. However, an incorrect skeleton can be detected when unnecessarily many joints exist because the limitations of the human body are not considered. A skeleton is a collection of joints, and a 3D skeleton represents the human skeletal joints in 3D coordinates. As a person moves over time, his/her action changes based on changes in the human skeleton. Therefore, the skeleton must be captured continuously at a regular interval. The skeleton sequence generated for an action has a 3D data format. This 3D data can be converted into a 2D image by projecting the coordinates into an RGB space. Fig. 3 shows the process of converting a skeleton sequence into an image by projecting a skeleton into an RGB space. Here, the skeleton sequence is expressed as 3D data (J × D × T), where J denotes the number of joints representing the human body skeleton, and T is the temporal dimension denoting the number of skeleton frames over time. To convert the skeleton sequence into an image, the dimension of the joint coordinates (D) is substituted with the temporal dimension (T). The substitution process produces a single-color image (J × T × 3) when the dimension of D is three. Normalizing this color image per channel and linearly transforming the image size yield a skeleton image. A pre-

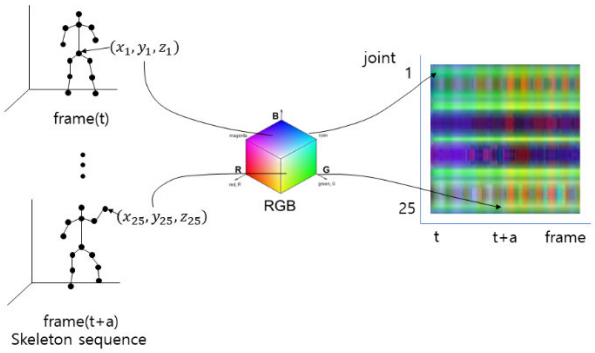


FIGURE 3. Process of converting skeleton sequence into image by projecting it onto RGB space.

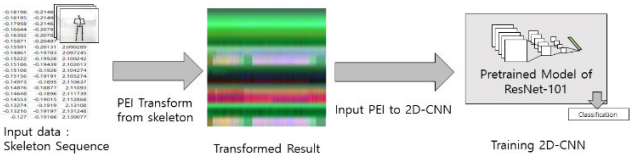


FIGURE 4. 2D-CNN using converted PEI as input.

trained 2D-CNN was designed to use three RGB channels as an input for image recognition. Therefore, the skeleton sequence can be used for the pre-trained 2D-CNN by converting it to a PEI. Moreover, both temporal and spatial features can be considered using only a 2D filter by converting the skeleton sequence into a PEI. Fig. 4 shows the process of a skeleton sequence being converted to a PEI and then used as an input to the 2D-CNN. Algorithm 1 depicts the pseudocode of ensemble three-stream RGB-S deep neural network.

III. ENSEMBLE RGB-S DEEP NEURAL NETWORK

In general, an ensemble method is a method for deriving better results by combining the results of several models, which have been trained individually with a single goal. Each model focuses on its assigned features without being affected by another model's input data. In addition, the analysis strategy of data can be diversified by having different neural networks for individual models. A more robust classifier is obtained by combining these models of diverse inputs and analysis strategies. The ensemble method used in this study adds or multiplies the output score values of the neural network models using the late fusion method.

The behavior recognition data typically comprise two types of data: RGB videos and skeleton sequences. These two data types exhibit different characteristics; therefore, combining them appropriately can yield a better synergy effect. Three RGB video-based behavior recognition models, RGB 2D-CNN + LSTM, RGB 3D-CNN, and skeleton PEI-CNN, were combined in this study.

For the PEI-CNN model, the performances of four types of PEI were analyzed. The four types of PEI comprised the following: (i) created based on the original skeleton data (Type 1), (ii) created by assigning rotation to the skeleton

Algorithm 1 Ensemble Three-Stream RGB-S Deep Neural Network

1. Load data:
   - Signal1 ← RGB Video
   - Signal2 ← Skeleton
2. Load pretrained model:
   - Model1 ← GoogLeNet
   - Model2 ← ResNet-101
   - Model3 ← R3D-18
3. Stage 1 of RGB 2D-CNN + LSTM
   - Signal1 = I = $\{I_1, \ldots, I_N\}$, N = video length
   - $F_i = Model1_{FeatureLayer}(I_i)$
   - $G = [F_1, \ldots, F_p]$, p = sequencelength
   - Training: LSTM(G_{train})
   - $Prob_1$ = LSTM(G_{test})
4. Stage 2 of Skeleton PEI-based CNN
   - Signal2 = S = $\{S_1, \ldots, S_N\}$, N = video length
   - H = PEI(S)
   - Training: Model2(H_{train})
   - $Prob_2$ = Model2(H_{test})
5. Stage 3 of RGB 3D-CNN
   - Signal1 = I = $\{I_1, \ldots, I_N\}$, N = video length
   - U = $[I_1, \ldots, I_p]$, p = sequencelength
   - Training: Model3(U_{train})
   - $Prob_3$ = Model3(U_{test})
6. Stage 4 of Ensemble Network
   - $Prob_{Fusion1} = Prob_1 + Prob_2 + Prob_3$
   - $Prob_{Fusion2} = Prob_1 \times Prob_2 \times Prob_3$
   - $Pred_{Fusion1} = \max(Prob_{Fusion1})$
   - $Pred_{Fusion2} = \max(Prob_{Fusion2})$
7.

(Type 2), (iii) created by adding joint points (Type 3), and (iv) created by assigning rotation to the skeleton and adding joint points (Type 4).

The RGB-S based three-stream ensemble deep neural network provides better performance by considering both the circumstance and color information of the RGB video and the human skeleton information of the skeleton sequence for the final behavior recognition. The RGB-S based three-stream ensemble deep neural network reflects the temporal and spatial characteristics of the actions in the video, and it includes the characteristics of the 2D sequence image, 3D video, and skeleton.

Fig. 5 shows the RGB-S based three-stream ensemble deep neural network for video-based behavior recognition.

IV. EXPERIMENTAL RESULTS

For a natural interaction between humans and robots and for providing customized services under intelligent home service robotic environments, the behavior recognition performance of the ensemble three-stream deep neural network was verified via the large-scale ETRI-Activity3D behavior

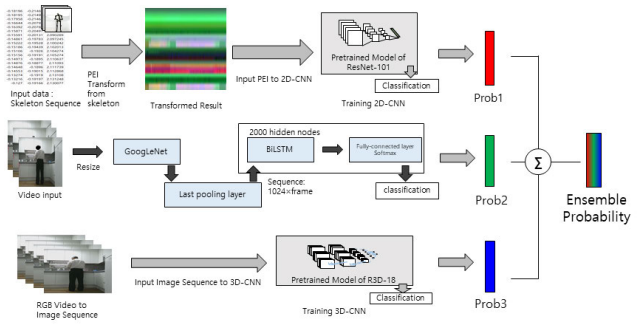


FIGURE 5. RGB-S based three-stream ensemble deep neural network.

recognition database. This section describes the experiment and evaluation method, as well as the experimental results and consideration.

A. ETRI-Activity3D DATABASE

To evaluate the performance of the behavior recognition model proposed herein, the large-scale ETRI-Activity3D dataset built by ETRI was used [5]. These data comprised 112,620 samples, and they were collected from 50 elderly people and 50 adults. The elderly group comprised 17 men and 33 women. Their ages ranged between 64 and 88 years, and the average age was 77.1 years. The adult group comprised 25 men and 25 women. Their ages ranged between 21 and 29 years, and the average age was 23.6 years. The participants performed 55 types of actions that reflected daily activities in the living room, kitchen, and bedroom of the residential apartment, where intelligent home service robots moved around. Their actions were recorded using the Kinect v2 version robot camera. The 55 types of actions were defined by observing frequently occurring daily activities of adults and elderly people. The detailed information can be viewed or downloaded from a public website. Considering home service situations, four Kinect sensors were installed at 70 and at 120 cm heights. The actions were recorded from eight directions, and the distance between the imaging device and the participant was between 1.5 and 3.5 m. The collected data presented a 1920 × 1080 resolution for a color image, and 512 × 424 resolution for a depth image. The skeleton information comprised 25 positions of the joints in a 3D space, and the frame rate of the data was 20. Fig. 6 shows sample images of the 55 action types. Table 1 shows 55 action types in the ETRI-Activity 3D dataset as an example.

Fig. 7 shows examples of action data in the ETRI-Activity3D dataset. For each action type, one person performs the action two to three times by either changing the location inside the house (living room, bedroom, kitchen, etc.) or the direction faced by the person. To diversity the data, 100 people performed the same action, and four to eight cameras, depending on the space condition, were used to record each person's action. Approximately 2050 data points were obtained for each action type, and the average number of data point for each person was 20.5. Owing to the enormous

TABLE 1. 55 Action Types in Etri-activity3d.

1	eating food with a fork	29	hanging out laundry
2	pouring water into a cup	30	looking around for something
3	taking medicine	31	using a remote control
4	drinking water	32	reading a book
5	placing (taking) food in (from) the fridge	33	reading a newspaper
6	trimming vegetables	34	handwriting
7	peeling fruit	35	talking on the phone
8	using a gas stove	36	playing with a mobile phone
9	cutting vegetable on a cutting board	37	using a computer
10	brushing teeth	38	smoking
11	washing hands	39	clapping
12	washing face	40	rubbing face with hands
13	wiping face with a towel	41	doing freehand exercise
14	putting on cosmetics	42	doing neck roll exercise
15	putting on lipstick	43	massaging a shoulder oneself
16	brushing hair	44	taking a bow
17	blow drying hair	45	talking to each other
18	putting on a jacket	46	handshaking
19	taking off a jacket	47	hugging each other
20	putting (taking) on (off) shoes	48	fighting each other
21	putting(taking) on(off) glasses	49	waving a hand
22	washing the dishes	50	flapping a hand up and down
23	vacuumming the floor	51	pointing with a finger
24	scrubbing the floor with a rag	52	opening the door and walking in
25	wipping off the dining table	53	fallen on the floor
26	rubbing up furniture	54	sitting(standing) up
27	spreading(folding) bedding	55	lying down
28	washing a towel by hands		

total data size, the image resolution was reduced to one-fifth, i.e., 384 × 216.

B. PERFORMANCE EVALUATION

Based on the experimental standards used in previous studies, the performance was evaluated using the cross-subject (CS) and cross-age (CA) of the dataset. For the CS, the 50 elderly people were assigned numbers 1 through 50, whereas the 50 adults were assigned numbers 51 through 100 based on ETRI-Activity3D. Numbers that were not multiples of three were categorized as learning data, whereas numbers that were multiples of three were categorized as verification data. Hence, the learning data comprised the data of 67 people, including both the adults and elderly. The verification data comprised the data of 33 people, also including both the adults and elderly. For the CA, data of the elderly and adults were separated, and it was composed of elderly learning data, elderly verification data, adult learning data, and adult verification data. Similar to the CS, the elderly people were assigned numbers 1 through 50, and the adults were assigned numbers 51 through 100 for the CA. Numbers that we're not multiples of three were categorized as learning data, whereas numbers that were multiples of three were categorized as verification data. The domains were separated at the boundary between the elderly and adults, between numbers 50 and 51 [24]. Fig. 8 shows the composition of ETRI-Activity3's CS, and Fig. 9 shows the composition of ETRI-Activity3's CA.

FIGURE 6. Sample images of the 55 action types for ETRI-Activity3D.

The trend of the behavior recognition dataset shows that the data were typically composed of RGB images, depth images, and skeletons. These data can be recorded simultaneously because sensors that obtain such data are built into a single device. Diverse sensor information can improve the accuracy as it provides more information for data analysis. In this study, RGB video images and skeleton sequences were used. As for the depth image, many images were obtained inaccurately. Hence, depth images were excluded to increase the recognition rate.

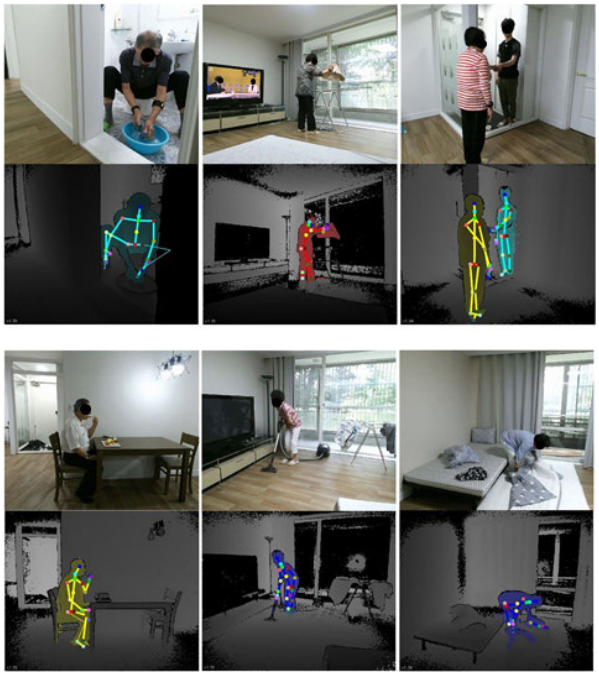


FIGURE 7. Examples of action data in ETRI-Activity3D.

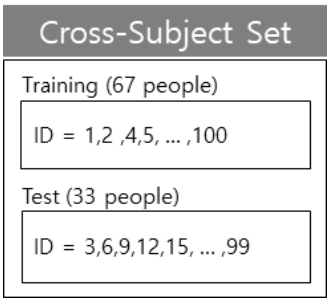


FIGURE 8. Division of cross-subject dataset.

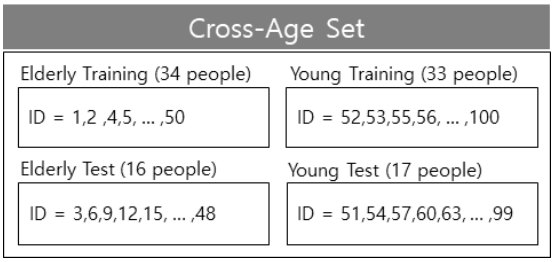


FIGURE 9. Division of cross-age dataset.

The behavior recognition model was evaluated based on accuracy, which was calculated by dividing the number of correct classifications (*Nc*) by the sum of the number of correct classifications (*Nc*) and the number of incorrect classifications (*Nw*). Equation 1 shows the formula for calculating the accuracy.

$$\text{accuracy} = \frac{N_c}{N_c + N_w} \tag{1}$$

TABLE 2. Performance comparison of proposed and existing methods (cs).

methods	recognition rate (%)
IndRNN [25]	73.90
Beyond joints [26]	79.10
SK-CNN [27]	83.60
ST-GCN [28]	86.80
Motif ST-GCN [29]	89.90
Ensem-NN [30]	83.00
MANs [31]	82.40
HCN [32]	88.00
FSA-CNN [24]	90.60
RGB 2D-CNN+LSTM	49.47
RGB 3D-CNN	79.20
Skeleton PEI-2D-CNN	86.09
Ensemble DNN (Fusion1)	91.02
Ensemble DNN (Fusion2)	93.20

C. EXPERIMENTAL RESULTS

The equipment used in the experiment comprised an Intel(R) Xeon(R) Gold 5120 2.2 GHz central processing unit, an NVIDIA Tesla V100-SXM2-32GB graphics processing unit (GPU), 180 GB of random access memory, and a 64-bit Windows Server 2016 operating system.

In the experiment, a recognizer was designed using the 2D-CNN + LSTM method and a 3D-CNN for the RGB videos in the three-stream configuration. For the skeleton sequence, the PEI-CNN method was used to design a recognizer.

A pre-trained GoogLeNet model was used as the 2D-CNN to implement the RGB 2D-CNN + LSTM model. Pre-trained models are open models that have been trained and proven to exhibit a good structure. The characteristics of neural networks allow them to be designed with diverse components and depth structure, but the performance varies significantly based on its structure. Hence, many trial-and-error processes are required. However, pre-trained models can shorten this process and can be applied promptly. Furthermore, they do not require tens to several hundreds of layers to be defined individually. The initial weights that had been learned through ImageNet were used to extract feature vectors in the last pooling layer. A feature vector was generated for each frame of a single video, and these feature vectors were used to create a sequence. Sequence data with a length of 400 were not included to exclude anomalous data from learning. The LSTM was designed using a BiLSTM layer comprising 2000 nodes and a dropout that reduced overfitting. The Adam optimization method was used, and the learning parameters set were as follows: a dropout rate of 0.5, an initial learning rate of 0.0001, the number of learning epochs of 30, a mini-batch size of 16, a learning rate decay period of 5, and a learning rate decay rate of 0.2. The experimental result showed that the 2D-CNN + LSTM yielded a recognition performance of 49.47%, as shown in Table 2.

The 3D-CNN is a CNN with 3D input data. The filters that are designed inside for feature extraction are 3D as well. Both spatial and temporal features are calculated in a

3D convolution operation; therefore, the 3D-CNN can learn the sequence data better than the 2D-CNN. As a 3D image feature extractor and classifier, the 3D-CNN used in this study utilized a pre-trained R3D-18 model. The Adam optimization was used and the parameters set were as follows: the number of learning epochs of 50, a learning rate of 0.001, weight decay of 0.00005, and a mini-batch size of 100. The experimental result showed that the recognition performance of the RGB 3D-CNN was 79.20%, as shown in Table 2.

ResNet-101 was used as the 2D-CNN in the implementation of the skeleton (PEI-2D-CNN) model. Furthermore, the initial weights learned through ImageNet were used to perform transfer learning. The 3D skeleton sequence was converted to images using the PEI method. Subsequently, those images were used as input to the 2D-CNN for classification. In addition, the performances of the four types of PEIs were analyzed.

In this study, the four types of PEI were generated based on the skeleton data. The four types of PEIs comprised the following: (i) created based on the original skeleton data (Type 1), (ii) created by assigning rotation to the skeleton (Type 2), (iii) created by adding joint points (Type 3), and (iv) created by assigning rotation to the skeleton and adding joint points (Type 4). In addition, the converted images were measured $224 \times 224 \times 4$ and presented in RGB format. Learning was performed using the Adam optimization method, a mini-batch size of 30, an initial learning rate of 0.0001, the number of learning epochs of 20, a learning rate decay period of 5, and a learning rate decay rate of 0.2.

The recognition performances of the four types of PEIs were 84.95%, 85.88%, 86.09%, and 85.20%. Type 3, which demonstrated the best performance, was selected in this study. As shown in Table 2, the recognition rate for skeleton PEI-2D-CNN was 86.09%.

Table 2 shows the performances of the previously investigated methods, individual methods of the three stream, and the ensemble DNN method. The ensemble DNN uses two types of late fusion methods. Fusion1 (summation) and Fusion2 (multiplication) indicate the method types used.

For the existing methods shown in Table 2 for comparison, the default values of the open-source models and the Adam optimization method were used. The learning rate was randomly set from 1/3 times to 3 times for each iteration until the weight decay. After the weight decay, the learning rate was reduced by 1/3 from 0.001 to 0.000001. The learning was performed variously by randomly setting the mini-batch size from 1 time to 1/4 times for each iteration based on the maximum GPU memory [24].

As shown in Table 2, the proposed method, i.e., the Fusion2 type of the ensemble DNN, outperformed the existing methods and individual stream methods by 2.6% to 20%, respectively.

Fig. 10 shows the recognition performance based on the learning of the PEI-2D-CNN having the skeleton as the input, whereas Fig. 11 shows a part of the PEI converted from the skeleton sequence. Fig. 12 shows the recognition performance based on the learning of the 3D-CNN having an RGB video as the input.

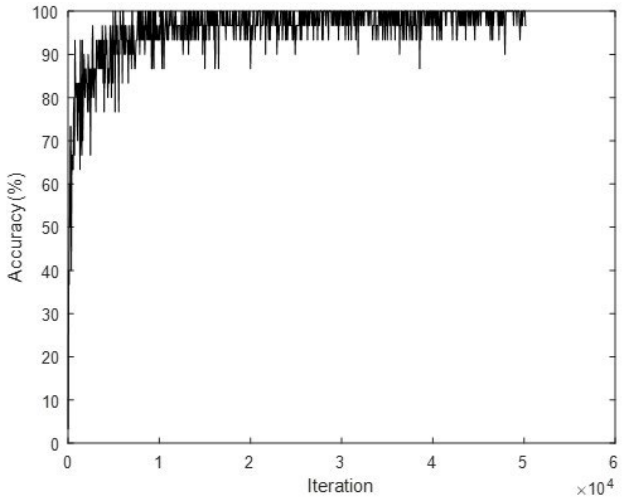


FIGURE 10. Recognition performance based on learning of skeleton PEI-2D-CNN.

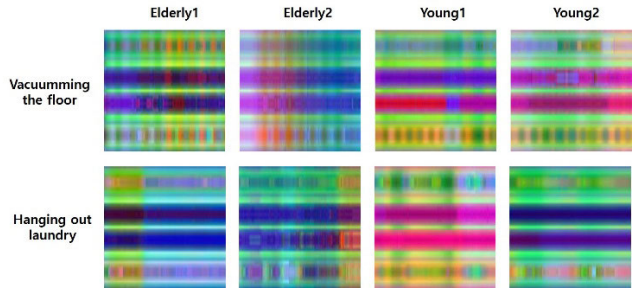


FIGURE 11. Part of PEI converted from skeleton sequence.

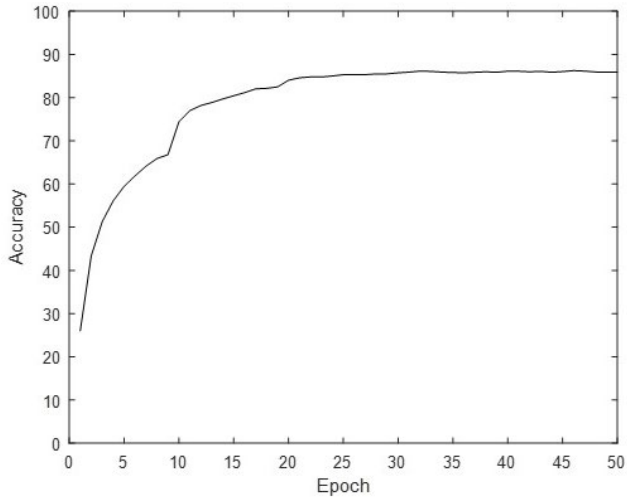


FIGURE 12. Performance recognition based on learning of RGB 3D-CNN.

CA performance verification was performed by segregating the participants into elderly and adult groups to evaluate the performance of the three-stream RGB-S ensemble DNN

TABLE 3. Performance comparison of proposed and existing methods (ca).

methods \ verification groups	Elderly	Adults
FSA-CNN (Case1) [24]	87.7	69.0
FSA-CNN (Case2) [24]	74.9	85.0
Skeleton PEI-2D-CNN(case1)	85.11	66.80
Skeleton PEI-2D-CNN(case2)	70.37	83.27
RGB 2D-CNN+LSTM (case1)	49.72	29.06
RGB 2D-CNN+LSTM (case2)	28.18	42.73
RGB 3D-CNN (Case1)	77.25	55.18
RGB 3D-CNN (Case2)	47.16	83.47
Ensemble DNN-Fusion1 (Case 1)	90.78	68.96
Ensemble DNN-Fusion1 (Case 2)	70.96	87.16
Ensemble DNN-Fusion2 (Case 1)	92.81	73.56
Ensemble DNN-Fusion2 (Case 2)	75.08	90.37

model and analyze the domain difference. This performance verification was performed separately from the previous CS experiment. The learning options and implementation environment were set to be the same as that of the CS experiment. Table 3 shows the ensemble accuracy of the behavior recognition based on the RGB and skeleton. The learning of the elderly and group datasets was labeled as Case1 and Case2, respectively. In the Case1 experiment, the ensemble DNN-Fusion1 achieved performances of 92.81% and 73.56%, whereas previously researched methods achieved performances of 87.7% and 69%. Hence, it was confirmed that the proposed ensemble DNN-Fusion1 method yielded better performances than the existing methods. Likewise, in the Case2 experiment, the existing methods achieved performances of 74.9% and 85%, whereas the proposed ensemble DNN-Fusion1 model achieved excellent performances of 75.08% and 90.37%.

D. DISCUSSION

The contributions of this paper can be summarized as follows. First, a large-scale database built under the robot environments was essential for the study of human-robot interaction to care for the elderly in intelligent home service robots. However, there was a very lack of data to recognize human behavior in human daily life. In particular, the research was difficult in research because a database specialized for the elderly was not constructed. To solve this difficulty, we constructed the large-scale ETRI-Activity3D dataset to evaluate the performance of human behavior recognition.

Second, we presented the method for recognizing behaviors in videos based on the ensemble RGB-S deep neural networks. This approach reflected the spatiotemporal features of the behaviors in videos, including all characteristics of the 2D sequence images, 3D videos, and skeleton sequences. The experimental results revealed that the presented approach showed good performance in comparison to the previous methods. From these contributions, this paper is expected to obtain an important starting point in research on the commercialization of robots for solving social problems in the age of aging and is considered to be of academic value.

However, although the proposed method offered better performance than other methods, its limitation was that it lacks in explaining the unique behavioral characteristics of the elderly and adults based on deep learning. An explainable AI method needs to be performed to solve this problem as future research. For readers, ETRI has released the ETRI-Activity3D database and software necessary for research on human care robots that help the elderly in their daily life [33].

V. CONCLUSION

A video-based behavior recognition method based on the ensemble RGB-S deep neural network was proposed herein. This method combines RGB videos and skeleton features from a behavior recognition database built under intelligent home service robot environments. The ETRI-Activity3D database, a large-scale video recognition database constructed by ETRI, was used to evaluate the performance of the proposed model. The recognition performance was 93.2% for the CS experiment. In the CA experiment, the proposed model achieved better performances than the existing models. The proposed behavior recognition technology is an HRI technology for intelligent home service robots. It is anticipated that this technology will be useful for future applications. Using deep learning technology, we plan to perform studies regarding methods for explaining unique behavioral characteristics of the elderly and adults in the future. Furthermore, we plan to develop a behavior recognition solution specialized for the behaviors of elderly people. This solution would be applied to eldercare robots.

APPENDIX

Pretrained models are famous preconfigured and weight-adjusted deep learning models that are already reviewed on performance scientifically. There are some pre-trained models such as AlexNet, GoogLeNet, ResNet, and DenseNet for 2D-CNN and I3D and R3D for 3D-CNN. The 2D-CNN and 3D-CNN uses 2D filters and 3D filters for convolution and pooling, respectively. The GoogLeNet applied Inception modules avoiding gradient vanishing to stack many layers deeply. The Inception modules are designed to process input data in parallel with varying filter size from 1 to 5, then the results are concatenated for feeding next layer [17]. The ResNet applied skip connection to add input to output for maintaining input information and prevent gradient vanishing [18]. So, the neural networks can have hundreds of layers successfully. The I3D [20] and R3D [21] are 3D-CNNs based on GoogLeNet and ResNet, respectfully. The 2D filter for convolution can only calculates spatial features and

the 3D filter can calculate spatiotemporal features. Commonly, 3D-CNN is used for recognizing time series data. The R3D-18 consists of 4 types of residual blocks and repeats each type twice. At part of end, averaging pooling and fully-connected layers are followed.

LSTM takes time series data and finds rule along temporal space. It deals with input and previous output recursively and has variable input and output lengths. To memorize previous information, it consists of input, forget, and output gates. The flow of LSTM calculation is as follows:

$$i_t = sigmoid(x_t w^t + h_{t-1} u^i) \tag{2}$$
$$f_t = sigmoid(x_t w^f + h_{t-1} u^f) \tag{3}$$
$$o_t = sigmoid(x_t w^o + h_{t-1} u^o) \tag{4}$$
$$\tilde{c_t} = \tanh(x_t w^{\tilde{c}} + h_{t-1} u^{\tilde{c}}) \tag{5}$$
$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c_t} \tag{6}$$
$$h_t = \tanh(c_t) \times o_t \tag{7}$$

where, w and u are weights multiplied to input and output, respectively. i_t, f_t, and o_t are input, forget, and output gates, respectively. c_t and h_t are cell state and output, respectively [34].

REFERENCES

[1] D. Kim, J. Lee, Y. Yoon, W. H. Yun, H. S. Yoon, and J. Kim, "HRIDemon: A framework for recognition of human location, identify and behavior in human-robot interaction," in *Proc. 9th Int. Conf. Ubiquitous Robots Ambient Intell. (URAI)*, Daejeon, South Korea, 2012, pp. 533–534.

[2] T. D. Le, D. T. Huynh, and H. V. Pham, "Efficient human-robot interaction using deep learning with mask R-CNN: Detection, recognition, tracking and segmentation," in *Proc. 15th Int. Conf. Control, Automat., Robot. Vis. (ICARCV)*, 2018, pp. 162–167.

[3] C. Zhu and W. Sheng, "Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 41, no. 3, pp. 569–573, May 2011.

[4] A. Manzi, A. Moschetti, R. Limosani, L. Fiorini, and F. Cavallo, "Enhancing activity recognition of self-localized robot through depth camera and wearable sensors," *IEEE Sensors J.*, vol. 18, no. 22, pp. 9324–9331, Nov. 2018.

[5] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "ETRI-Activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, Oct. 2020, pp. 10990–10997.

[6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," 2014, *arXiv:1406.2199*. [Online]. Available: http://arxiv.org/abs/1406.2199

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

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

[9] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodel fusion of deep neural networks for video classification," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 978–987.

[10] K. Zhang and W. Ling, "Joint motion information extraction and human behavior recognition in video based on deep learning," *IEEE Sensors J.*, vol. 20, no. 20, pp. 11919–11926, Oct. 2020.

[11] Z. Wang, K. Jiang, Y. Hou, W. Dou, C. Zhang, Z. Huang, and Y. Guo, "A survey on human behavior recognition using channel state information," *IEEE Access*, vol. 7, pp. 155986–156024, 2019.

[12] A. Gong, C. Chen, and M. Peng, "Human interaction recognition based on deep learning and HMM," *IEEE Access*, vol. 7, pp. 161123–161130, 2019.

[13] J. Liu, C. Wang, Y. Gong, and H. Xue, "Deep fully connected model for collective activity recognition," *IEEE Access*, vol. 7, pp. 104308–104314, 2019.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Ohio, CO, USA, Jun. 2014, pp. 1725–1732.

[15] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," 2015, *arXiv:1503.08909*. [Online]. Available: http://arxiv.org/abs/1503.08909

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, USA, 2015, pp. 1–9.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: http://arxiv.org/abs/1512.03385

[19] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*. [Online]. Available: http://arxiv.org/abs/1608.06993

[20] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," 2017, *arXiv:1705.07750*. [Online]. Available: http://arxiv.org/abs/1705.07750

[21] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2018, pp. 6450–6459.

[22] *Github*. Accessed: Nov. 10, 2020. [Online]. Available: https://github.com/jfzhang95/pytorch-video-recognition

[23] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2018, pp. 1159–1168.

[24] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "ETRI-activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, New York, NY, USA, Oct. 2020, pp. 10990–10997.

[25] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (IndRNN): Building a longer and deeper RNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5457–5466.

[26] H. Wang and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4382–4394, Sep. 2018.

[27] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2017, pp. 597–600.

[28] S. Yan, X. Yuanjun, and L. Dahua, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.

[29] Y. H. Wen, L. Gao, H. Fu, F. L. Zhang, and S. Xia, "Graph CNNs with motif and variable temporal block for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8989–8996.

[30] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, "Ensemble one-dimensional convolution neural networks for skeleton-based action recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 1044–1048, Jul. 2018.

[31] C. Xie, C. Li, B. Zhang, C. Chen, J. Han, C. Zou, and J. Liu, "Memory attention networks for skeleton-based action recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1639–1645.

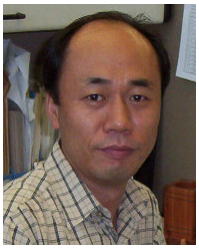
[32] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 786–792.

[33] *ETRI-Activity3D*. Accessed: Sep. 5, 2020. [Online]. Available: https://ai4robot.github.io/etri-activity3d-en/

[34] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, vol. 2, 2014, pp. 3104–3112.

YEONG-HYEON BYEON received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Control and Instrumentation Engineering, Chosun University, Gwangju, South Korea, in 2013, 2015, and 2021, respectively. He is currently a Postdoctoral Fellowship with IT Institute, Chosun University. His research interests include behavior recognition, explainable artificial intelligence, pedestrian detection, and deep learning.

DOHYUNG KIM received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Computer Engineering, Pusan National University, Busan, South Korea, in 2000, 2002, and 2009, respectively. He is currently a Principal Researcher with the Human–Robot Interaction Research Section, Intelligent Robotics Research Division, Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. He is also a Professor with the Department of Computer Software, University of Science and Technology, Daejeon. His research interests include human–robot interaction, computer vision, and machine learning.



JAEYEON LEE received the Ph.D. degree from Tokai University, Japan, in 1996. He has been a Research Scientist with the Electronics and Telecommunications Research Institute (ETRI), since 1986. His research interests include robotics, pattern recognition, computer vision, and biometric security.

KEUN-CHANG KWAK (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Electronics Engineering, Chungbuk National University, Cheongju, South Korea, in 1996, 1998, and 2002, respectively.

From 2003 to 2005, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada. He was also a Senior Researcher with the Human–Robot Interaction Team, Intelligent Robot Division, Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, from 2005 to 2007. He was also a Visiting Professor with the Department of Computer Science, California State University, Fullerton, USA, from 2014 to 2015. He is currently a Professor with the Department of Electronics Engineering, Chosun University, Gwangju, South Korea. His research interests include computational intelligence, human–robot interaction, and biometrics.