S & M 3826

Human Activity Recognition System Based on Continuous Learning with Human Skeleton Information

Wenbang Dou,* Aulia Saputra Azhar, Weihong Chin, and Naoyuki Kubota

Graduate School of Systems Design, Tokyo Metropolitan University,
6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan

(Received September 25, 2024; accepted October 8, 2024)

Keywords: human skeleton model, human activity recognition, continuous learning

In recent years, as the demographic profile of society continues to shift towards an aging population, there has been a concomitant shortage of caregivers, leading to an increase in the demand for elderly care. The accurate assessment of the health status of the elderly and the provision of appropriate care necessitate the timely recognition and analysis of human activities. To address this challenge, we propose a continuous human activity recognition system that generates a 3D human skeleton model, utilizes joint angles to perform daily life activity recognition, and infer similarities in movements across various body parts. The proposed system generates a 3D human skeleton model using depth information obtained from multiple range-based depth cameras and extracts human joint angles on the basis of this model. Moreover, it utilizes time-series joint angle data to continuously recognize actions and estimate the similarity of movements across various body parts. To validate the efficacy of the proposed system, comprehensive verification experiments were conducted using real-world data.

1. Introduction

In recent years, the rapid advancement of technology has had a significant impact on a number of areas, including manufacturing, social infrastructure, and healthcare. The effect of technology on these areas has been considerable, affecting aspects such as the quality of life, convenience, and the way healthcare is delivered. The advent of surgical and communication robots has resulted in a profound transformation in the landscape of medical treatment and patient care. Furthermore, the global phenomenon of aging populations, prevalent in developed countries worldwide, has resulted in a significant increase in the demand for health maintenance and caregiving services for the elderly.

To address these challenges, there is currently active research in the field of artificial intelligence technology, with a particular focus on the development of wearable devices, nursing-care robots, and intelligent agents for the detection of human conditions. Kaburagi *et al.* proposed a system that employs the growing neural gas to extract text features, with the objective of providing support for care records and case management in caregiving facilities.(1) Shao *et al.*

*Corresponding author: e-mail: dou-wenbang@ed.tmu.ac.jp
https://doi.org/10.18494/SAM5374

developed a monitoring system that employs high-precision vibration sensors to quantify the periodic human activities of elderly individuals living alone.(2) The system was designed from the informationally structured space perspective and with a time-delay neural network that can analyze vibration data.(2) Abrar *et al.* concentrated on cardiovascular diseases and put forth a multi-agent-based hypertension risk prediction system that incorporates a Gaussian mixture model and an online infinite echo state Gaussian process.(3) Besari *et al.* addressed the rehabilitation of post-stroke patients with visual impairments, proposing a cyber–physical–social system that employs perception-based egocentric vision for classifying four primary activities (wonder, reach, grasp, and release) during hand–object interactions in the grasping task.(4)

The application of these technologies enables the continuous and effective monitoring of the health status of the elderly and patients, thereby facilitating the early detection of a disease and the promotion of rehabilitation. Nevertheless, the necessity of regulating human circadian rhythms, evaluating health status, and administering prompt treatment to the elderly and patients without unduly disrupting their daily lives makes it imperative to estimate human physical states from daily life movements. Furthermore, the provision of detailed information to the elderly and caregivers in caregiving situations requires the gathering of specific data regarding individual body parts based on daily life activities. In this study, we propose a continuous human activity recognition system that measures human physical capabilities without being intrusive, monitors human activities over time in a naturalistic manner, learns and recognizes human daily activities, and infers problem areas using the similarity of movements among various body parts.

2. Related Works

In the context of body monitoring, the tracking of the human skeleton model and the extraction of human posture features are of paramount importance. The tracking of the human skeleton model has a variety of applications, including in the fields of security, rehabilitation, and entertainment. At present, the recognition of human skeleton models is occasionally accomplished through the use of electromagnetic sensors,(5) inertial sensors,(6) and other analogous devices. Inertial measurement units (IMUs) can detect the inertial motion of human movements, thereby enabling the acquisition of high-precision data even during activities with high degrees of freedom.(7) However, the use of IMU sensors presents certain challenges, including limited battery life and the inability to directly track positions.(8) Consequently, the generation of human skeleton models using wearable devices places significant physical and psychological burdens on users, thereby rendering continuous daily-life attachment of such devices challenging. An alternative methodology entails the attachment of multiple markers to disparate regions of the body and the utilization of multiple high-frequency cameras for the generation of the human skeleton model. Nevertheless, this methodology presents challenges associated with camera interference.(9)

To facilitate the identification of human skeleton models, markerless motion capture systems that rely solely on camera technology have been proposed as a means of developing automated methods for their generation.(10) Cao *et al.* developed the OpenPose multi-skeleton tracking module, which generates real-time human skeleton models from a single camera.(11) Similarly, Pauzi *et al.* developed the Mediapipe Blazepose, a camera-based system for the generation of human skeleton models.(12) However, the recognition of obscured body parts using a single camera remains a significant challenge, and the development of solutions to address this limitation is essential for the practical application of these methods in human monitoring in daily life.

A multitude of technologies have been put forth in previous reports on human activity recognition. Conventional human activity recognition systems entail the acquisition of data from wearable sensors or images for the purpose of recognizing and learning about multiple actions. Gholamiangonabadi and Grolinger proposed a model for human activity recognition from wearable sensor data using convolutional neural networks (CNNs) and signal analysis.(13) In a further development of this field, Yu *et al.* introduced densely 3D-long short term memory (D3D-LSTM), which combines 3D-CNN and LSTM to enable the real-time recognition of prolonged actions in complex environments based on RGB-D data.(14) Coppola *et al.* addressed the context of ambient assisted living with a method based on qualitative trajectories to represent human actions, which are learned and classified using hidden Markov models.(15) Although these human activity recognition systems establish the structure of the model for learning, the accumulation of previously learned data is essential to address the catastrophic forgetting problem that arises when learning new activities. In daily life, where human activity is subject to considerable variation due to individual differences, the data accumulated each time the system undergoes relearning becomes extensive, leading to high computational costs for retraining. Therefore, in the context of daily life, a model that can continuously learn with flexibility to recognize human activities is of paramount importance.

Moreover, there has been a growing emphasis on research utilizing human skeleton models to enhance the accuracy of human activity recognition in traditional activity recognition systems.(16) In a related vein, Li *et al.* put forth a human activity recognition system that employs RGB-D data.(17) In this system, the human skeleton model is transformed into joint distance maps and identified through a CNN approach.(17) In a comparative study, Nguyen *et al.* evaluated the performance of the recurrent neural network, CNN, graph convolutional network, and hybrid–deep neural network using input from a 3D human skeleton model.(18) Dou *et al.* proposed declarative memory recurrent neural model (DM-RNM) with self-organizing adaptive recurrent incremental network (SOARIN), a continuous learning model for multiple interpretations of hand gestures.(19) However, these activity recognition systems frequently perceive the entire human body as a singular entity, thereby impeding the capability to analyze distinctions in body parts for a given movement due to individual variations. To gain a precise understanding of human body states in everyday life, it is vital to examine the discrepancies in the same activity resulting from individual differences while performing activity recognition.

The continuous human activity recognition system proposed in this study comprises two components: human posture detection and human activity recognition. These components are designed to address the aforementioned issues. (1) Human posture detection: In the context of monitoring human posture, the system addresses occlusion issues through the utilization of data from multiple RGBD cameras (Azure Kinect). The system combines a 3D human skeleton model on the basis of reliability of the skeleton models generated from each camera. This multi-viewpoint approach facilitates the overcoming of occlusion challenges and enhances the accuracy of the 3D human skeleton model. (2) Human activity recognition: In the context of human activity recognition, the system employs a methodology that entails the calculation of key joint angles from an anatomical perspective, with the objective of estimating human posture. The system extracts spatiotemporal features from the temporal data of human activity, thereby enabling learning. To prepare for the recognition of unknown activities, the human activity recognition model employs a continuous learning model, which allows for learning and recognition over time. Furthermore, the system infers the similarity of movement among body parts to present issues with the movements of each body part during various actions, thereby aiding users in understanding the challenges associated with each action.

3. System Design

The objective of this study is to develop a continuous human activity recognition system that can track human posture with consistent reliability and continuously learn and recognize human actions. The aim is to develop a system that can perform human activity recognition and body state estimation in seamlessly within the context of daily life, without requiring the user's awareness or input. The proposed system's architectural design is illustrated in Fig. 1. The proposed system is structured with a two-layer architectural configuration, comprising posture detection and activity recognition. The system employs depth data from multiple Azure cameras in a manner that does not infringe upon human privacy to generate skeleton models. To address the issue of occlusion, the skeleton models generated from each camera are combined on the basis of reliability to construct a 3D human skeleton model. To facilitate the straightforward acquisition and recognition of human activity characteristics in the subsequent model, the

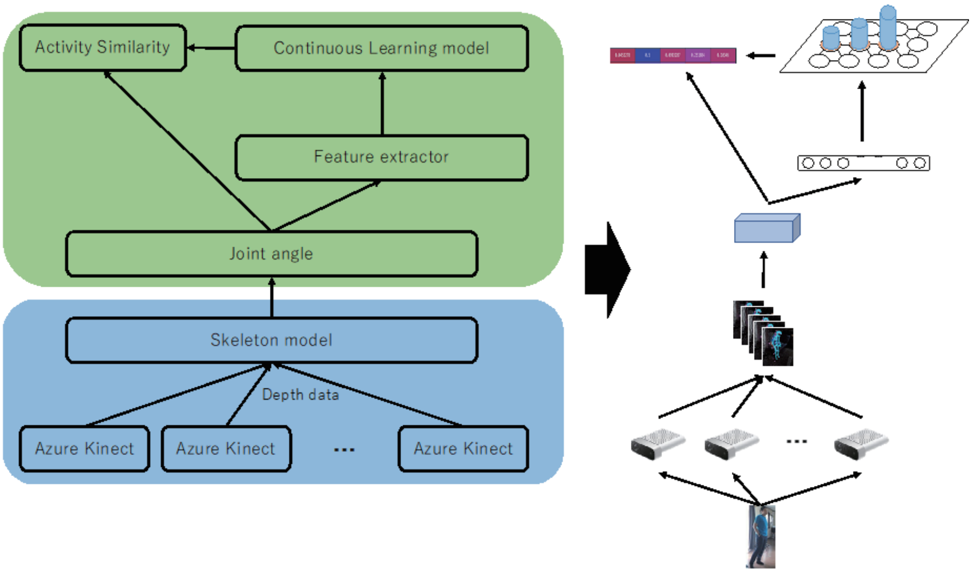

Fig. 1. (Color online) Architecture of proposed system.

essential joint angles are calculated from an anatomical perspective using the constructed 3D human skeleton model. To facilitate the recognition of human activities, a feature extractor is employed to extract the spatiotemporal characteristics of human actions from the human joint angles. Concurrently, the system identifies discrepancies in the manner in which each body part is utilized during a given activity by inferring the degree of similarity of movements among body parts and estimating the challenges associated with the movement of each body part during routine activities. Moreover, when the system is deployed in real-world scenarios, the continuous emergence of novel activities is affected by human habits and bodily states. To accommodate the recognition of previously unidentified activities, the system employs the random weight convolutional-growing memory network (RWC-GMN) continuous learning model, which enables the automatic adjustment of the model size without the need for presetting.(20)

3.1 Human posture monitoring

In the context of Azure's human body skeleton recognition, passive infrared and depth information are employed as inputs. Infrared data is employed as input to a pretrained CNN utilizing the ONNX Runtime, which was trained on a human body skeleton dataset. The CNN generates 32 2D joint positions. The system employs the 2D frame position data of joint angles as depth information, transforming them into 3D coordinates that are contingent upon the camera's position. Subsequently, as illustrated in Fig. 2, the integration of multiple 3D skeleton datasets is proposed as a means of enhancing accuracy and addressing occlusion issues.(21) The 3D joint positions of multiple camera modules are transformed on the basis of six-dimensional information (x-position, y-position, z-position, pitch, roll, and yaw). Subsequently, pelvic and head joint positions are calculated, and movement similarity detection is conducted to identify

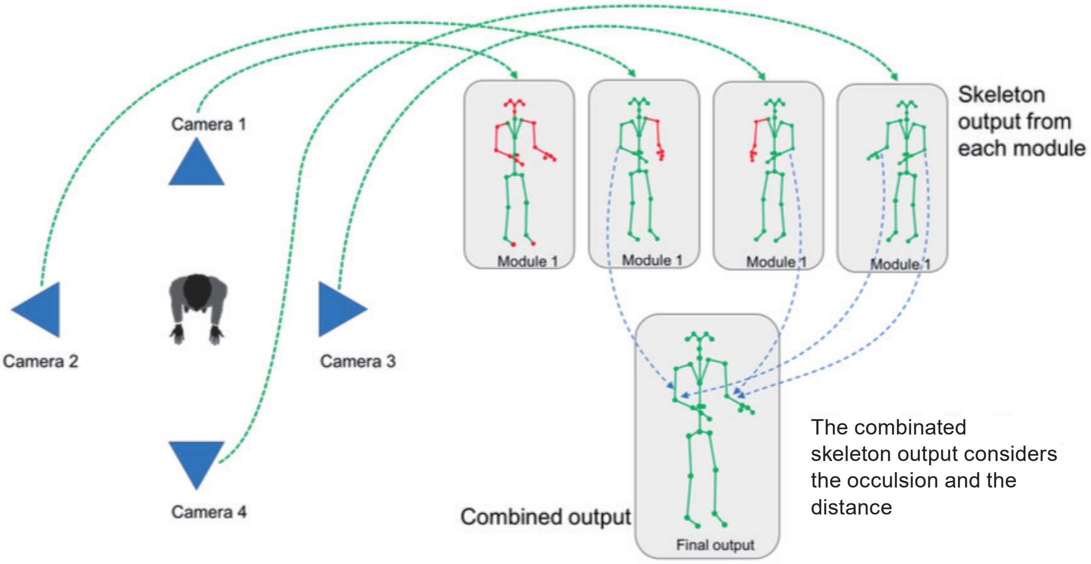
Camera 1
Camera 2
Camera 3
Camera 4
Module 1
Module 1
Module 1
Module 1
Skeleton output from each module
Final output
Combined output
The combinated skeleton output considers the occlusion and the distance

Fig. 2. (Color online) Illustration of multiple skeleton datasets for recognition.

the skeleton representations. Subsequently, a combinatorial process is employed to combine multiple skeleton datasets on the basis of a confidence analysis. The skeleton model shown in Fig. 3 is used to categorize joints into six groups: torso, head, left upper limb, right upper limb, left lower limb, and right lower limb. The combined joint position [Eq. (3)] is calculated on the basis of the confidence analysis $C(t)$ of the joint groups shown in Eqs. (1) and (2).

$$C_k^i(t) = \sum_j^{N_j} \left(c_{j,k}^i + o_{j,k}^i \right), \tag{1}$$

$$C(t) = \sum_i^{N_i} \frac{C_k^i(t)}{\sum_k^{N_k} C_k^i(t)}, \tag{2}$$

$$J(t) = J(t-1) + C(t), \tag{3}$$

where $J(t)$ is the vector position of the joint at time t and $C(t)$ is the confidence level of the joint group at time t. The total confidence level is calculated on the basis of the single confidence level $c_{j,k}^i$ of the j-th joint of the k-th joint group at time t of the i-th skeletal model and the shielding $o_{j,k}^i$

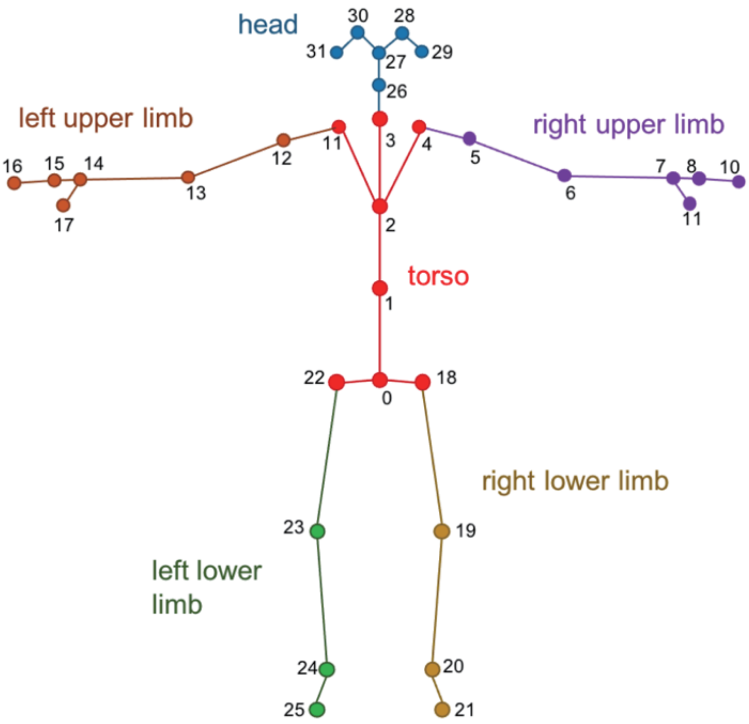

Fig. 3. (Color online) Division of the joint group and the joint ID.

of the j-th joint of the k-th joint group at time t of the i-th skeletal model. The parameters N_i, N_k, and N_j represent the number of skeletal models, the joint groups, and the number of joints detected at the same location, respectively.

$$o = \begin{cases} 1 & \text{if } \theta > 0 \\ e^d & \text{otherwise} \end{cases} \tag{4}$$

The shielding analysis is conducted using Eq. (4) on the basis of Fig. 4, where θ is the deviation of the joint positions in different joint groups and α is the deviation of the depth of the joint positions in different joint groups. The parameters θ and α are calculated using Eqs. (5) and (6), respectively, and the degree of occlusion is calculated using Eq. (7).

$$\theta = \left(\left(J - P_k\right) \cdot \hat{n}\right) \cdot n, \tag{5}$$

$$\alpha = \left(J - P_k\right) - \theta, \tag{6}$$

$$d = \frac{\alpha}{\theta} - \beta, \tag{7}$$

where P_k is the k-th camera position and $\hat{n}$ is the vector unit of the difference $J - P_k$ between the joint position J and the camera position. β is the gradient threshold.

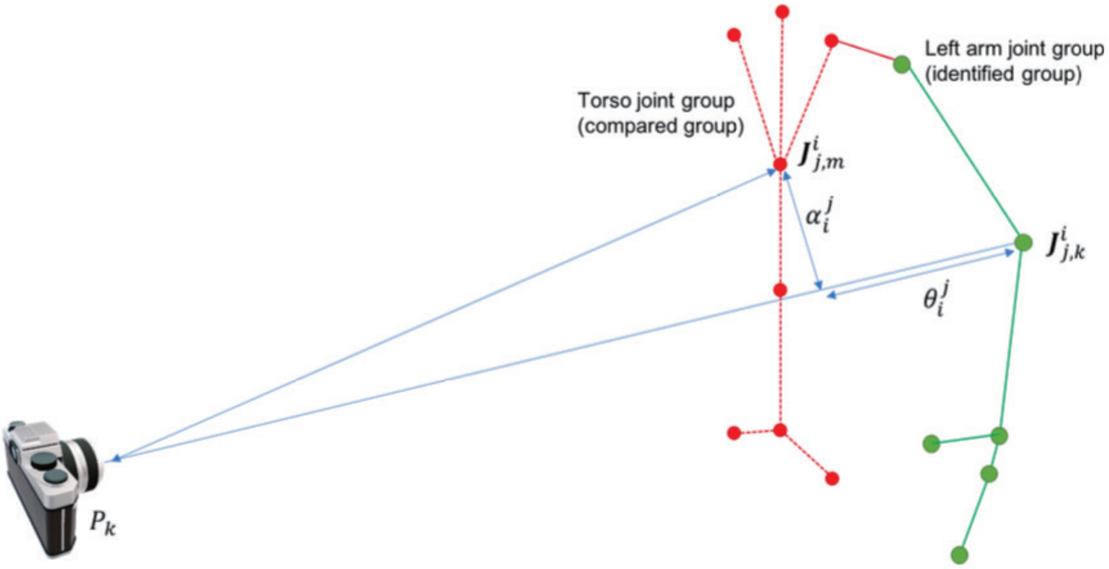

Fig. 4. (Color online) Confidence analysis based on joint position.

3.2 Human activity recognition

3.2.1 Joint angle estimation

To recognize human activity and analyze the differences among body parts in performing the same action, angle information for each major body joint is estimated from the human skeletal model and used as a feature of human activity. As illustrated in Fig. 3, the output of the human skeletal model yields 32 3D joint positions, with the X, Y, and Z values representing relative values from the 3D Cartesian coordinate system. The calculation of human joint angles is primarily based on three rotational axes: the coronal, the anterior–posterior, and vertical axes, depending on the direction of rotation. In this example, the one-DOF elbow joint and three-DOF shoulder joint of the right hemisphere will be used as examples for the calculation. Consequently, identical formula may be employed to calculate the analogous joints of the left half of the body using the mirror image positions of the body pose information. Table 1 also presents the major joint angles of the entire body.

3.2.1.1 Elbow joint angle calculation

Since the elbow joint operates in extension and flexion with one-DOF around the coronal axis, the calculation of joint angles is simple. As shown in Eqs. (8) and (9), the elbow joint (joint ID: P_6) is used to calculate a vector representation consisting of the shoulder joint (joint ID: P_5) and wrist joint (joint ID: P_7), and the angle between the vectors is calculated.

Table 1
List of major human joint angles.

Joint	Angle number	Axis
Neck	angle1	coronal axis
	angle2	anterior-posterior axis
	angle3	vertical axis
Body	angle4	coronal axis
	angle5	anterior-posterior axis
	angle6	vertical axis
Right shoulder	angle7	coronal axis
	angle8	anterior-posterior axis
	angle9	vertical axis
Right elbow	angle10	coronal axis
Left shoulder	angle11	coronal axis
	angle12	anterior-posterior axis
	angle13	vertical axis
Left elbow	angle14	coronal axis
Right leg	angle15	coronal axis
	angle16	anterior-posterior axis
	angle17	vertical axis
Right knee	angle18	coronal axis
Left leg	angle19	coronal axis
	angle20	anterior-posterior axis
	angle21	coronal axis
Left knee	angle22	anterior-posterior axis

$$\begin{cases} v_{(6,5)} = P_5 - P_6, \\ v_{(6,7)} = P_7 - P_6, \end{cases} \tag{8}$$

$$A_{elbow} = \cos^{-1} \frac{v_{(6,5)} \cdot v_{(6,7)}}{\left\| v_{(6,5)} \right\| \cdot \left\| v_{(6,7)} \right\|}, \tag{9}$$

where $v_{(6,5)}$ is the 3D vector representation of the elbow and shoulder joints, and A_{elbow} is the joint angle of the elbow joint.

3.2.1.2 Shoulder joint angle calculation

Three angles are calculated for the joint angle of the shoulder joint: the shoulder joint elevation plane angle $A_{s\_e\_p}$ around the anterior–posterior axis, the shoulder joint elevation angle A_{s_e} around the coronal axis, and the shoulder joint rotation angle A_{s_r} around the vertical axis. To calculate the plane angle of shoulder joint elevation and the angle of shoulder joint rotation in 3D space, a torso plane that can be used as a reference is constructed using the pelvis joint (joint ID: P_0), left shoulder joint (joint ID: P_{12}), and right shoulder joint (joint ID: P_5), and then the normal vector $\hat{n}_{body}$ of the torso plane is calculated as in Eqs. (10) and (11).

$$\begin{cases} v_{(0,5)} = P_5 - P_0 \\ v_{(0,12)} = P_{12} - P_0 \end{cases} \tag{10}$$

$$\hat{n}_{body} = v_{(0,5)} \times v_{(0,12)} \tag{11}$$

Next, the projection point P_{proj_e} of the elbow joint (joint ID: P_6) in the torso plane is calculated from Eq. (12). The shoulder joint elevation plane angle $A_{s\_e\_p}$ is calculated on the basis of Eqs. 12 and 13 using the projection point P_{proj_e} of the elbow, the shoulder joint (joint ID: P_5), and the chest joint (joint ID: P_2).

$$P_{proj\_e} = P_6 - \frac{v_{(5,6)} \cdot \hat{n}_{body}}{\left\| \hat{n}_{body} \right\|^2} \cdot \hat{n}_{body} \tag{12}$$

$$\begin{cases} v_{(5,2)} = P_2 - P_5 \\ v_{\left(5,p_{r_e}\right)} = P_{proj\_e} - P_5 \end{cases} \tag{13}$$

$$A_{s\_e\_p} = \cos^{-1} \frac{v_{(5,2)} \cdot v_{\left(5,p_{r_e}\right)}}{\left\|v_{(5,2)}\right\| \cdot \left\|v_{\left(5,p_{r_e}\right)}\right\|} \tag{14}$$

The calculation of the shoulder joint elevation angle A_{s_e} is output from the elbow joint (joint ID: P_6), shoulder joint (joint ID: P_5), and projection point P_{proj_e} of the elbow as shown in Eqs. (15) and (16).

$$v_{(5,6)} = P_6 - P_5 \tag{15}$$

$$A_{s_e} = \cos^{-1} \frac{v_{(5,6)} \cdot v_{\left(5,p_{r_e}\right)}}{\left\|v_{(5,6)}\right\| \cdot \left\|v_{(5,p_{r_e})}\right\|} \tag{16}$$

The shoulder joint rotation angle A_{s_r} is calculated using the angle between the normal vector $\hat{n}_{elbow}$ of the plane composed of the shoulder joint (joint ID: P_5), elbow joint (joint ID: P_6), and wrist joint (joint ID: P_7) and the normal vector $\hat{n}_{shoulder}$ of the plane composed of the elbow joint (joint ID:), shoulder joint (joint ID: P_5), and projection point P_{proj_e} at the elbow [Eqs. (17) and (18)].

$$\begin{cases} \hat{n}_{elbow} = v_{(6,7)} \times v_{(6,5)} \\ \hat{n}_{shoulder} = v_{\left(5,p_{r_e}\right)} \times v_{(5,6)} \end{cases} \tag{17}$$

$$A_{s\_r} = \cos^{-1} \frac{\hat{n}_{elbow} \cdot n_{shoulder}}{\left\|\hat{n}_{elbow}\right\| \cdot \left\|n_{shoulder}\right\|} \tag{18}$$

3.2.2 Human activity recognition model

The objective of this study was to develop a continuous human activity recognition system that can learn and recognize human activity from time series data of human major joint angles. These angles represent the features of human activity and can be used to distinguish between different activities and body parts. The system was designed to continuously learn human activity and detect differences in body parts from human activity by using the range of motion (ROM) rank, as illustrated in Fig. 5. In this paper, we use the ROM of human joints as defined by the Japanese Ministry of Health, Labor and Welfare as ROM.

To accurately discern human activity features, the pre-extracted time series data of joint angles are divided into two parts: long-term features that exhibit long-term spatiotemporal characteristics and short-term features that represent the features of movements in a brief period. Subsequently, a fixed random weight 3DCNN is employed as a feature extractor to transform the

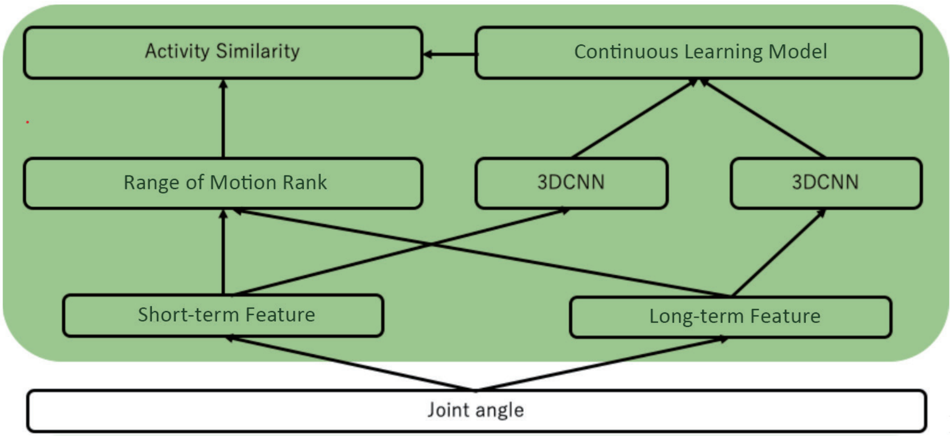

Fig. 5. (Color online) Human activity recognition system.

data into features that can be trained into a continuous learning model, and two types of spatiotemporal feature are further extracted.

Fixing random weights is a machine learning technique in which the weights of a neural network are randomly initialized at the outset of the learning process and then fixed during the learning process itself. This differs from the process of updating the weights of the network through the conventional method of backpropagation. The rationale behind the significance of fixed random weights is that they can diminish variability within the learning process, thereby enhancing the stability of the acquired policies. In particular, when weights are fixed, the learning algorithm can explore the solution space in greater depth without overfitting the training data. This enhances the generalization performance on novel data and prevents the network from becoming constrained in a local minimum (a partially optimal solution that is not a global minimum). Another significant benefit of fixed random weights is that they can enable continuous learning. Continuous learning is the capacity to adapt to new data over time without forgetting previously acquired knowledge. The utilization of fixed random weights permits the continuation of learning from new data whilst retaining a degree of plasticity and adaptability, thereby preventing the catastrophic forgetting of previously acquired knowledge.

To enable continuous learning of new data, we have proposed a continuous learning model, the growing memory network (GMN), which comprises a self-organizing topological network that emulates human episodic memory. Figure 6 depicts the GMN architectural configuration, and the learning process is as follows:

On the basis of the sensory input x, the network first generates two episode nodes and updates the long-term memory weight. For time t = 1, each element of the long-term memory weight becomes as follows.

$$\mathbb{G}_{max} = \mathbb{G}_{min} = x(1) \tag{19}$$

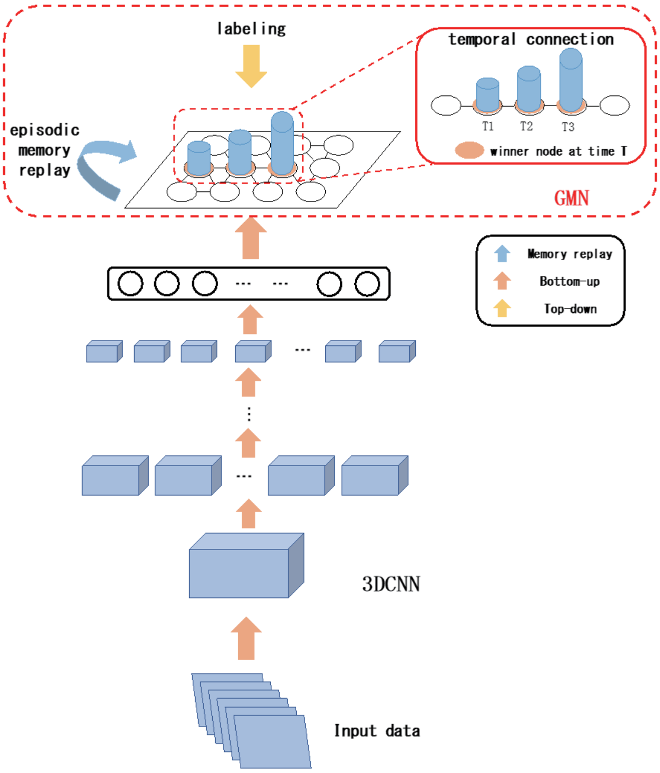

Fig. 6. (Color online) Architecture of the continuous learning model GMN.

Afterwards, the long-term memory weight is updated as follows.

$$\mathbb{G}_{max} \leftarrow \mathbb{G}_{max} + \beta \cdot \left(\max_{i \in (0,N)} \left(\mathbb{G}_{max,i}, x_i(t) \right) - \mathbb{G}_{max,i} \right) \tag{20}$$

$$\mathbb{G}_{min} \leftarrow \mathbb{G}_{min} + \beta \cdot \left(\min_{i \in (0,N)} \left(\mathbb{G}_{min,i}, x_i(t) \right) - \mathbb{G}_{min,i} \right) \tag{21}$$

Here, i represents the index of each element in long-term memory weight. max(a, b) and min(a, b) represent the maximum and minimum values, respectively, for each element.

Owing to the characteristics of the long-term memory weight, the system is less affected by the setting of vigilance parameters and can continue learning without normalizing input data. Next, each episode node in the network is composed of a weight vector w_j. Using Eqs. (22) and (23), the network selects the winner node for the current input data $x(t)$ for the next learning process.

$$b = \arg\min\left(T_j(t) \right) \tag{22}$$

$$T_j(t)=\frac{\left\|x(t)-w_j\right\|^2}{\left\|\mathbb{G}_{max}-\mathbb{G}_{min}\right\|^2} \tag{23}$$

Then, the activation value of the winning node *J* is calculated as

$$a_b(t)=exp(-T_b). \tag{24}$$

If the activation value $a_b(t)$ is less than the threshold a_T that was initially set, a new node *N* is added to the network with new weights as follows.

$$w_N=0.5\cdot\left(x(t)+w_b\right) \tag{25}$$

To connect the selected best winner node *b* and the second winner node, it is necessary to generate a new edge. If $a_b(t)$ is greater than a_T, the best winner node *b* can represent the input $x(t)$. As a result, the best winner node *b* and its neighboring node *n* are updated using the input $x(t)$ as follows.

$$w_j\leftarrow w_j+\gamma_j\cdot r_j\cdot\left(x(t)-w_j\right) \tag{26}$$

If there is no edge between the best winner node *b* and the second winner node, a new edge is generated to connect them. For each learning iteration, the age counter of each edge is increased by one. The age counter of the edge between the best winner node and the second winner node is initialized to zero. Nodes without edges, nodes with a habituation counter greater than the threshold, and edges with an age counter greater than the threshold are removed from the network. Additionally, each episode node has a regularity counter r_j indicating the firing strength over time, ranging from [0, 1]. The regularity counter of the newly formed episode node is initialized to $r_j = 1$. The regularity counter of the best winner node and its neighbor nodes are decayed for each learning iteration using the following formula.

$$r_j\leftarrow r_j+\tau_j\cdot\lambda\cdot\left(1-r_j\right)-\tau_j \tag{27}$$

As a result, the regularity counter of a node can express the relevance and importance of the information stored in that node. The regularity counter [Eq. (27)] indicates the regularity value of a node that is triggered over time depending on the learning input. If the age counter of the edge is greater than the age threshold, these independent nodes and edges are removed from the network. To prevent the removal of useful edges generated at the beginning of learning, we remove nodes according to the criteria introduced in a previous study(22) as follows.

$$v = \mu(H) + \sigma(H) \tag{28}$$

Here, H is a vector representation of the regularity counter of all nodes in the network, μ is the mean function, and σ is the standard deviation. Nodes with a regularity counter greater than the threshold are removed.

The new episode node is connected to the network only if $b_J(t) < \rho_b$ and $r_J < \rho_r$. If both the activation and regularity thresholds are met, the node is updated using Eq. (26). In the GMN network, a succession of events creates an episode that recalls distinctive prior experiences and episodes related to one another to simulate episodic memory properties. The activation patterns of episode nodes in the network are learned using temporal connections. Temporal connections represent the order in which the activated nodes occurred during the learning stage. If the best winner node b is activated at time t and other nodes were activated at time $t-1$, the temporal connection between them is reinforced.

$$P_{(b(t),b(t-1))} \leftarrow P_{(b(t),b(t-1))} + 1 \tag{29}$$

Therefore, for each episode node m of the encoded time series, the next node g can be obtained by selecting the maximum value of P as follows.

$$g = \arg\max P_{(m,n)} \tag{30}$$

Here, n is the neighbor node of m. The activation time sequence of episode nodes can be restored without requiring input data.

On the basis of previous research, the episode memory network is applied to utilize the spatiotemporal connectivity of existing nodes to replay meaningful temporal data. The episode memory network can replay temporal data when no sensory input is provided. For example, the best winner node b is activated by the input in episode memory network nodes. The next temporal connection can be generated by selecting the node with the highest activation value P. For each node j, the replay memory of length $K+1$ is calculated as

$$U_j = \left\langle w_{u(0)}, w_{u(1)}, \ldots, w_{u(K)} \right\rangle, \tag{31}$$

$$u(i) = \arg\max P_{(j,u(i-1))}. \tag{32}$$

Here, $P(i, j)$ is a matrix of temporal connections for all episode with $u(0) = j$. It is possible to automatically generate a sequence of memories and replay them in the network without storing previously learned data, by establishing temporal connections for existing episode nodes in the network.

During the learning phase, class label l can be assigned to each node on the basis of input data. L classes generate l class labels. In this labeling method, the frequency of each label in the network is stored in $V(j, l)$. Using this, each node j holds a distribution counter that maintains the frequency of the specific label assigned to it. A new node N is created, and the label ζ associated with the input data $x(t)$ is determined. The matrix V is extended by one row, and $V(N, \zeta) = 1$ and $V(N, l) = 0$ are initialized. When an existing winning node b is selected for weight updating, the V matrix is also updated as follows.

$$V(b,\zeta) \leftarrow V(b,\zeta) + \varphi^{+} \tag{33}$$

$$V(b,l) \leftarrow V(b,l) + \varphi^{-} \tag{34}$$

Note that φ^{+} must always be smaller than φ^{-}, and the label ζ belongs to class L. If the data label ζ does not exist in class L, a new column is added to V, and $V(b, \zeta) = 1$, $V(b, l) = 1$. If it does not match the label of the given input data, the matrix V is not updated. The selected label ζ_j for node j is calculated as

$$\zeta_j = label(j) \equiv \arg\max V(j,l). \tag{35}$$

Here, l is a label within class L. The advantage of this labeling method[23] is that it is not necessary to determine the class labels in advance. This method enables learning when the number of data classes is unknown.

To infer differences among body parts in human activity, the similarity of the ROM among body parts is calculated on the basis of joint angles. Specifically, the $Rank_i$ of each feature is calculated to L levels on the basis of the human joint ROM, and the similarity of ROM among body parts in human activity is calculated to infer differences.

$$Rank_i = \frac{L \cdot A_i}{Max\_Range_i - Min\_Range_i} \tag{36}$$

Here, A_i is the angle of each joint, and Max_Range_i and Min_Range_i are the maximum and minimum ROMs at each joint, respectively.

4. Experimental Results

To illustrate the efficacy of the proposed system in authentic settings, we conducted a comparative analysis with our previous continuous learning system, SOARIN,[24] and a conventional 3DCNN-based activity recognition system utilizing daily activities. In this experiment, models were trained on seven activities (Fig. 7) on the basis of the International Classification of Functioning (ICF) criteria: d415 (stand, sit), d430 (pick up), d450 (walk), d520

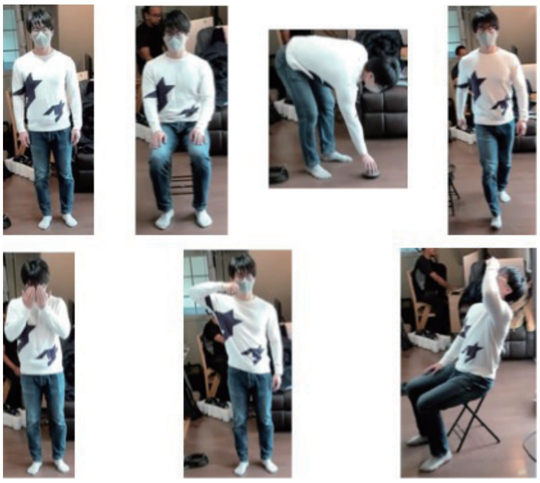

Fig. 7. (Color online) Daily activities based on ICF criteria.

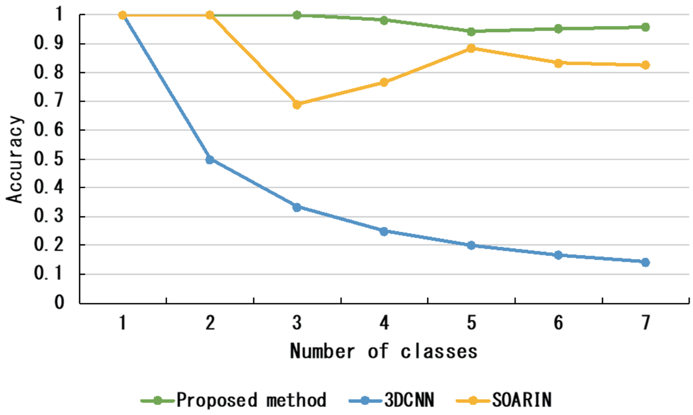

Fig. 8. (Color online) Experimental results comparing the proposed system, 3DCNN, and SOARIN for continuous learning.

(wash face, brush teeth), and d560 (drink). The 3D coordinates of the human skeleton model depicted in Fig. 3 were employed as data features for the purpose of learning. In the context of continuous learning, the data for each class was input sequentially and learned only once, with each model trained using the aforementioned activities. Upon the completion of the training phase for each class, the models were evaluated using all previously learned activities. The comparative results of the continuous learning for each model are presented in Fig. 8.

The results demonstrate that the proposed system exhibited continuous learning capabilities superior to those of the 3DCNN, with the capacity to retain and recall the acquired knowledge over time. Moreover, although the proposed system demonstrated a decline in the proportion of accurate responses as the number of learned activity classes increased, the recognition rate for each activity remained consistently higher than that of the conventional continuous learning model, SOARIN.

Table 2
Comparison results for each model on the classification task.

	Proposed system	3DCNN	SOARIN
Accuracy	**0.949**	0.929	0.827
Precision	**0.955**	0.952	0.883
Recall	**0.949**	0.929	0.827
F1	**0.949**	0.924	0.787
Continuous learning	○	×	○
Activities similarity	○	×	×

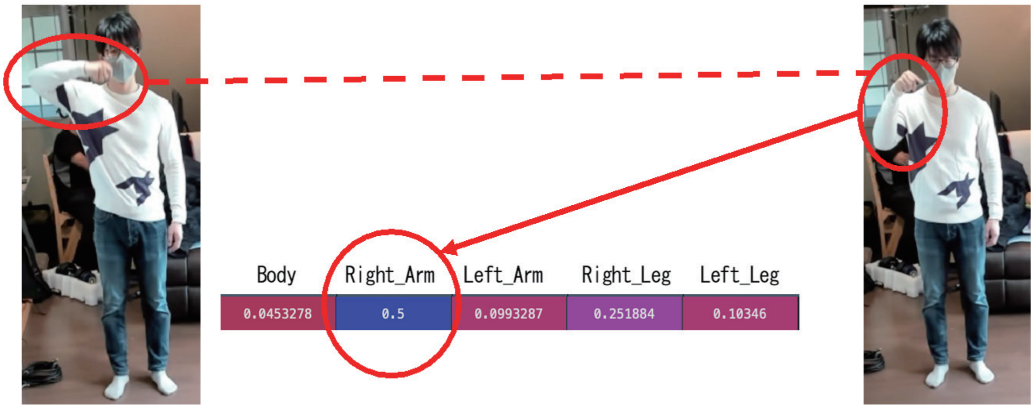

Fig. 9. (Color online) Difference estimation results of the same activity using the proposed system.

The classification task was also validated using all daily activities with the aforementioned model, and the resulting recognition rates are presented in Table 2. The results demonstrate that the proposed system exhibits a recognition performance in the classification task superior to that of other models. With regard to the same activity, the proposed system was capable of estimating the differences in the activity by calculating the similarity of movements among body parts based on the stored general knowledge, as illustrated in Fig. 9.

5. Conclusions

In this paper, we proposed a continuous human activity recognition system that learns and recognizes human daily activities over time. Furthermore, we utilized the similarity of activities among human body parts to infer and present the problematic part. A comparison of the proposed system with a conventional activity recognition system revealed that the former exhibits superior performance in continuously learning and recognizing new activities without forgetting previous knowledge. Moreover, the proposed system can estimate the discrepancies among body parts during the performance of a given activity and detecting the human body state with greater precision than the conventional action recognition system. As future work, the proposed system must be trained on high-dimensional activity features and validated on a large class of activities.

Acknowledgments

This work was (partially) supported by JST, [Moonshot R&D; Grant Number JPMJMS2034] and [the establishment of university fellowships towards the creation of science technology innovation; Grant Number JPMJFS2139], and TMU local 5G research support, and the Great Britain Sasakawa Foundation [Grant Number 6515].

References

1 R. Kaburagi, T. Obo, N. Kubota, and Y. Maeda: 2022 Joint 12th Int. Conf. Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS & ISIS, 2022) 1–5. https://doi.org/10.1109/SCISISIS55246.2022.10001914

2 S. Shao, N. Kubota, K. Hotta, and T. Sawayama: J. Adv. Comput. Intell. Intell. Inf. **25** (2021) 489.

3 S. Abrar, C. K. Loo, and N. Kubota: IEEE Access **9** (2021) 75090. https://doi.org/10.1109/ACCESS.2021.3074791

4 A. R. A. Besari, A. A. Saputra, W. H. Chin, N. Kubota, and Kurnianingsih: 2022 Int. Joint Conf. Neural Networks (IJCNN) (2022) 1–6. https://doi.org/10.1109/IJCNN55064.2022.9892903

5 R. W. Bohannon, S. Harrison, and J. Kinsella-Shaw: J. NeuroEng. Rehabil. **6** (2009) 1.

6 D. Roetenberg, H. Luinge, and P. Slycke: Xsens Motion Technol. BV Tech. Rep. **1** (2009) 1.

7 S. Sharma, S. Verma, M. Kumar, and L. Sharma: 2019 Int. Conf. Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (IEEE, 2019) 289–294.

8 M. Menolotto, D.-S. Komaris, S. Tedesco, B. O'Flynn, and M. Walsh: Sensors **20** (2020) 5687.

9 S. Park and S. Yoon: Sensors **21** (2021) 3667.

10 A. Sengupta, F. Jin, R. Zhang, and S. Cao: IEEE Sens. J. **20** (2020) 10032.

11 Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh: IEEE Trans. Pattern Analysis & Machine Intelligence **43** (2021) 172. https://doi.org/10.1109/TPAMI.2019.2929257

12 A. S. B. Pauzi, F. B. Mohd Nazri, S. Sani, A. M. Bataineh, M. N. Hisyam, M. H. Jaafar, M. N. Ab Wahab, and A. S. A. Mohamed: Advances in Visual Informatics: 7th Int. Visual Informatics Conf. (IVIC) (2021) 562–571.

13 D. Gholamiangonabadi and K. Grolinger: Appl. Intell. **53** (2023) 6041.

14 J. Yu, H. Gao, W. Yang, Y. Jiang, W. Chin, N. Kubota, and Z. Ju: IEEE Access **8** (2020) 43243. https://doi.org/10.1109/ACCESS.2020.2977856

15 C. Coppola, O. M. Mozos, and N. Bellotto: IEEE/RSJ IROS Workshop on Assistance and Service Robotics in a Human Environment (IEEE, 2015).

16 L. L. Presti and M. La Cascia: Pattern Recognit. **53** (2016) 130.

17 C. Li, Y. Hou, P. Wang, and W. Li: IEEE Signal Process Lett. **24** (2017) 624. https://doi.org/10.1109/LSP.2017.2678539.

18 H.-C. Nguyen, T.-H. Nguyen, R. Scherer, and V.-H. Le: Sensors **23** (2023) 5121. https://doi.org/10.3390/s23115121

19 W. B. Dou, W. H. Chin, and N. Kubota: 2020 Joint 11th Int. Conf. Soft Computing and Intelligent Systems and 21st Int. Symp. Advanced Intelligent Systems (SCIS-ISIS) (2020) 1–5. https://doi.org/10.1109/SCISISIS50064.2020.9322784

20 W. Dou, W. H. Chin, and N. Kubota: 2023 IEEE Int. Conf. Fuzzy Systems (FUZZ) (2023) 1–6. https://doi.org/10.1109/FUZZ52849.2023.10309751

21 A. A. Saputra, A. R. A. Besari, and N. Kubota: 2022 Int. Electronics Symposium (IES) (2022) 430–435. https://doi.org/10.1109/IES55876.2022.9888532

22 W. S. Liew, C. Kiong Loo, V. Gryshchuk, C. Weber, and S. Wermter: 2019 Int. Joint Conf. Neural Networks (IJCNN) (2019) 1–8.

23 I. G. Parisi, J. Tani, C. Weber, and S. Wermter: Front. Neurorob. **12** (2018) https://doi.org/10.48550/arXiv.1805.10966

24 W. H. Chin, N. Kubota, C. K. Loo, Z. Ju, and H. Liu: 2019 Int. Joint Conf. Neural Networks (IJCNN) (2019) 1–6. https://doi.org/10.1109/IJCNN.2019.8851919