Multistream Temporal Convolutional Network for Correct/Incorrect Patient Transfer Action Detection Using Body Sensor Network

Zhihang Zhong, Chingszu Lin, Masako Kanai-Pak, Jukai Maeda, Yasuko Kitajima, Mitsuhiro Nakamura, Noriaki Kuwahara, Taiki Ogata, and Jun Ota, *Member, IEEE*

***Abstract*—The development of body sensor networks (BSNs) with rich multimodal signals has enabled highly accurate fine-grained action detection, which is the cornerstone of many human–computer interaction applications. However, in the case of consecutive fine-grained actions, most existing wearable sensor-based detection methods are constrained by sliding windows because of their limited temporal receptive fields, and existing sequence-to-sequence detection methods cannot effectively leverage the potential of multimodal information of wearable sensors. Herein, to give multimodal signals full play in fine-grained action detection, we propose a novel temporal convolutional network by designing a channel attention-based multistream structure. We apply it to a promising application for correct and incorrect patient transfer nursing action detection. A data set is collected from a BSN on a patient when nurses perform patient transfer. Extensive experiments on our data set and public data set (C-MHAD) demonstrate that the proposed method is superior to the state-of-the-art methods, because it can strengthen the utilization of prediction features from the more convincing modal stream at each time frame. The source code is available at https://github.com/zzh-tech/Continuous-Action-Detection.**

***Index Terms*—Attention mechanism, body sensor network (BSN), deep learning, fine-grained action detection, multimodal signal.**

Manuscript received January 19, 2021; revised April 2, 2021; accepted April 17, 2021. Date of publication April 26, 2021; date of current version November 19, 2021. This work was supported in part by Grants-in-Aid for Scientific Research (KAKENHI) from the Japan Society for the Promotion of Science (JSPS) under Grant 20H04261, Grant 19H05730, and Grant 19H04154. *(Corresponding author: Zhihang Zhong.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Tokyo (UTokyo) and Tokyo Ariake University of Medical and Health Sciences (TAU).

Zhihang Zhong, Chingszu Lin, and Jun Ota are with the Research into Artifacts, Center of Engineering, University of Tokyo, Tokyo 113-8656, Japan (e-mail: zhong@is.s.u-tokyo.ac.jp; lin@race.t.u-tokyo.ac.jp; ota@race.t.u-tokyo.ac.jp).

Masako Kanai-Pak is with the Faculty of Nursing, Kanto Gakuin University, Yokohama 236-8501, Japan (e-mail: kanaipak@kanto-gakuin.ac.jp).

Jukai Maeda, Yasuko Kitajima, and Mitsuhiro Nakamura are with the Faculty of Nursing, Tokyo Ariake University of Medical and Health Sciences, Tokyo 135-0063, Japan (e-mail: jukai@tau.ac.jp; kitajima@tau.ac.jp; m-nakamura@tau.ac.jp).

Noriaki Kuwahara is with the Department of Advanced Fibro-Science, Kyoto Institute of Technology, Kyoto 606-8585, Japan (e-mail: nku-wahar@kit.ac.jp).

Taiki Ogata is with the Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo 152-8550, Japan (e-mail: ogata.t.af@m.titech.ac.jp).

Digital Object Identifier 10.1109/JIOT.2021.3075477

I. Introduction

With the rapid development of Internet of Things (IoT) in recent years, various kinds of sensors are now intimately embedded in and connected with our daily lives; these sensors record dynamic information about the human life in the form of multimodal signals. The detection of human actions from time-series signals recorded by sensors is an important topic concerning the IoT. In particular, fine-grained action detection is indispensable to IoT-based applications, such as skill training and health monitoring systems [1]–[3]. Fine-grained action detection, which is fairly challenging, requires the location and categorization of each occurring action based on the captured time-series signal [4]. Fine-grained actions are highly similar; their patterns only differ slightly at the signal level, which further increases the difficulty encountered in distinguishing between them [5]. Thus far, highly accurate fine-grained action detection has not been extensively studied; this detection (especially, its practical applications) requires further exploration.

The development of body sensor networks (BSNs) [6]–[8] has enabled highly accurate fine-grained action detection. A combination of wearable sensors, such as IMUs, placed on various parts of the body surface (i.e., BSN) can provide rich contextual activity information and easily capture the dynamic semantics of human activities in the temporal and spatial domains. In addition, for human action detection, BSNs are superior to camera-based sensor systems in the following aspects: 1) fewer privacy concerns [9]; 2) no complicated calibration processes, such as viewpoint fixation [10] and environmental constraints [11]; and 3) no viewpoint hindrance due to multiparticipant interaction. Therefore, BSNs are currently a favorable hardware platform for developing fine-grained action detection-based applications. However, the full utilization of multimodal information of BSNs for improving the recognition accuracy remains to be investigated.

In this study, we explored a new branch of smart hospitals for automatic nursing (healthcare) skill assessment. More specifically, we aim to develop an effective fine-grained action detection method, which can be applied to construct a self-help skill training system for assisting nursing learners in learning specific skills. It could make nursing skill training more efficient and convenient. This is important because the demand for

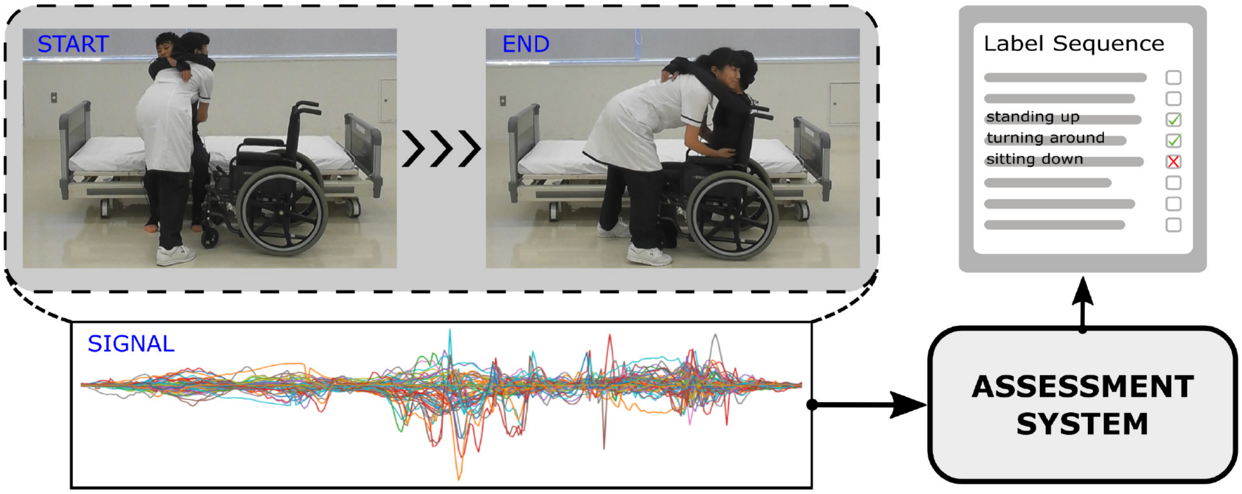


Fig. 1. Scheme for automatic nursing skill assessment. Inertial measurement unit (IMU) signals, including acceleration and rotational speed, are collected when a nurse performs patient transfer by deploying a BSN on the patient. The proposed detection method can generate a label sequence, including the action type and correctness by directly analyzing raw time-series signals.

qualified nurses is increasing owing to social issues [12] such as the aging population and increase in outbreaks of infectious diseases such as COVID-19 [13]. Owing to the high demands of healthcare environments, nurses are expected to provide effective and efficient care to patients, implying that nursing skills are essential. However, nursing education resources are limited owing to factors such as the high student–educator ratio. This hinders the acquisition of nursing skills and path to become qualified nurses because sufficient individualized supervision and feedback from traditional teaching methods are not provided to nursing learners during their training [14]. The previous study [15] has proven that showing whether a nursing action step is correct or not has an educational effect and can improve learning efficiency of the learners. Nursing learners can correct their wrong performance when training skills by themselves, according to the feedback on the correctness of action step. Therefore, developing a nursing skill assessment scheme that can automatically recognize and assess the correctness of nursing actions of trained nurses using fine-grained action detection will be beneficial.

According to previous studies [15], [16], the most basic requirement of a nursing skill assessment system is to inform the nursing learners about the correctness of each of their action steps; this feedback can allow the learners to correct their corresponding actions. Thus, the nursing skill assessment scheme proposed in this study primarily aims to recognize and assess the correctness of each performed nursing action based on the raw time-series signals of BSNs. There is no prior knowledge about the sequence of actions performed. The proposed scheme is illustrated in Fig. 1.

The most important requirement of a nursing skill assessment system is its reliability, which is determined by its nursing action recognition accuracy. An unreliable system may mislead nursing learners and reduce the effectiveness of learning. Action detection for nursing skill assessment is more challenging than that for activities of daily living because there are fine-grained action pairs in the case of nursing skill assessment, which reflect correct and incorrect ways of performing each nursing action step. Assisting a patient to stand up “with” (correct) and “without” (incorrect) bending their waist first is an example of a pair of fine-grained actions. To improve the action detection accuracy, many researchers have adopted deep learning methods, such as convolutional neural networks (CNNs) [11], [17], [18] and recurrent neural networks (RNNs) [19]–[21], and achieved satisfactory results. In the case of signals of consecutive actions, the most common pipeline of existing wearable sensor-based methods is the use of sliding windows (SWs) to segment the time-series signal, followed by single-action recognition for each SW. However, owing to the limited temporal receptive field, the SW size can easily affect the final performance; excessively large SWs will inevitably contain the information of adjacent actions, whereas significantly small SWs will contain only part of the target action signal [22], and in both the cases, the fine-grained action detection performance can be severely degraded. Recently, some sequence-to-sequence (seq2seq) action detection methods [23]–[25] have been proposed for video-based human action detection tasks. These seq2seq methods jointly generate predictions for each time frame based on the entire input time-series signal, without separate segmentation and recognition. These methods can obtain optimized temporal receptive fields for better performance. However, the existing seq2seq action detection methods do not consider the characteristics of multimodal wearable sensor signals, yielding suboptimal performance.

We believe that each separated modality or a combination of modalities may have advantages in a particular time frame when recognizing different actions of different individuals. Therefore, we propose a multistream temporal convolutional network (TCN) to realize both framewise and stepwise nursing action detection. A data set of nursing skill, called patient transfer, was collected for evaluation. We collected data samples of three consecutive actions of patient transfer by deploying BSNs on the body of a patient. Our data set includes the correct and incorrect methods for each action step of patient transfer. With the help of the proposed channel attention-based multistream structure, our method could better utilize the characteristics of multimodal wearable sensor signals when testing on the nursing skill data set.

The key contributions of this study are summarized as follows.

1) We propose a novel seq2seq action detection method to solve the automatic fine-grained nursing action detection problem using a multistream TCN.
2) The proposed channel attention-based multistream structure proved to be effective in utilizing raw multimodal wearable sensor signals.
3) Through extensive experiments on our nursing skill data set and publicly available data set (C-MHAD), we demonstrated the superiority of our method to the state-of-the-art human action detection methods.
4) To the best of our knowledge, this study was the first to collect a consecutive nursing action data set with correct/incorrect fine-grained action pairs using BSNs.

The remainder of this article is organized as follows. We briefly introduce related work in Section II. In Section III, we present the details of our nursing skill data set. Then, we explain the architecture of the proposed method in Section IV. The experimental configuration, results, and discussions are presented in Section V. Finally, we conclude the article and discuss future work in Section VI.

II. Related Work

In this section, the research on human action recognition (HAR) is presented. Then, human action detection methods are further discussed. Finally, the attention mechanism in deep learning is briefly introduced.

A. Human Action Recognition

Recent HAR methods can be roughly divided into three categories based on the sensors used: 1) wearable sensor-based methods; 2) camera-based methods [26], [27]; and 3) WiFi device-based methods [28], [29]. We focus on wearable sensor-based methods. Traditional wearable sensor-based HAR methods typically require two steps: 1) feature extraction and 2) classification. Experienced engineers need to manually decide the appropriate features from the time domain (e.g., variance mean, max, and min) or frequency domain (e.g., skewness, amplitude, DC, and energy) based on the characteristics of the target actions and their domain expertise. Various machine learning classifiers, such as support vector machines (SVMs) [30], decision trees (DTs) [31], and ensemble methods [32], are widely used to make action class predictions. In addition, hidden Markov models (HMMs) [33] are also used to model the intrinsic features and the continuous observed hidden state for the time-series signals of each action class. When new signals are fed, the output is determined based on the likelihood value of the HMM of each action class.

However, heuristic handcrafted features are often not informative enough and too shallow to help classical machine learning methods learn useful information for recognition. In addition, HMM is limited by the number of possible hidden states that it can have. Thus, an increasing number of deep learning methods have been applied for wearable sensor-based HAR tasks in recent years. The effectiveness of CNNs for wearable sensor-based HAR has been demonstrated in [11], [17], and [34]. For example, Jiang and Yin [17] first converted the time-series signals into activity images and then proposed a deep convolutional neural network (DCNN) that automatically learns the discriminative features for action recognition. Furthermore, owing to the strong ability of RNNs to capture the long-term temporal correlation between time-series signals, many variants of RNNs, such as DeepConvLSTM [35], DRNN [20], ResBidirLSTM [36], and MA-RNN [21], have been proposed to solve the HAR problem. Essentially, these methods utilize long short-term memory (LSTM) [37] to model wide-range dependencies of time-series signals recurrently. Bidirectional and stacked structures are used to enhance the recognition performance.

B. Human Action Detection

While all these HAR methods are based on segmented signals for individual action recognition, the task of human action detection is to recognize a series of actions based on unsegmented consecutive time-series signals. The mainstream scheme for human action detection is to segment consecutive signals through an SW and then recognize them successively with the help of the aforementioned HAR methods. This means that all the time frames within the same SW will be recognized as the same label. However, the SW-based scheme is limited by its temporal receptive field for single-action recognition. Determining the window size is often challenging as the appropriate window size depends on the type of corresponding action and even on the subject. A size either too large or too small will cause undesired conditions for recognition, introducing irrelevant temporal features or leading to information loss. To overcome this issue, some seq2seq detection methods have been proposed for realizing framewise recognition with optimized temporal receptive fields. For the prediction at each time frame, seq2seq methods automatically mine the relevant information from all the time-series signals based on their structural design. Deep Bi-LSTM [23] draws support from the transferred global hidden features of bidirectional LSTM [38]. Motivated by the WaveNet [39], dilated TCN [24], and MS-TCN [25] adopt stacked dilated convolutional layers to continuously improve the temporal receptive field for action detection. However, these seq2seq methods were originally designed for video signals; hence, they do not consider the characteristics of multimodal wearable sensor signals. We believe that each separate modality or a combination of modalities may have advantages at certain time frames when recording different actions of different individuals. Therefore, inspired by the structure of seq2seq methods, we designed a multistream TCN to fully utilize the different modalities of wearable sensor signals for achieving high detection accuracy.

C. Attention Mechanism

Allocating more attention toward more informative or convincing parts of signals is a wise strategy to improve the performance of deep learning models. The concept of attention mechanism [40], [41], a selective focusing rule, originated in the field of natural language processing. Speech and text signals are also a type of time-series data. Thus, the attention mechanism can be naturally applied to HAR tasks. For example, Zhong *et al.* [21] used an attention module to determine the time period in which discriminating features were located. Furthermore, Zhang *et al.* [18] attempted to utilize a multihead attention module to learn the relevance and importance of each feature produced by multihead CNNs. In this study, we applied the channel-attention mechanism [42] to our multistream TCN to give full play to the advantages of different wearable signal modalities.

III. Nursing Skill Assessment Data Set

In this section, we introduce our nursing skill assessment data set of a nursing skill named patient transfer. The main purpose of this data set is for developing data-driven nursing skill assessment schemes to alleviate the burden on nursing educators caused by high nursing student–educator ratio.

A. Target Nursing Skill

Nursing activities typically involve complicated and laborious physical work. To conserve energy, avoid injuries, and ensure the comfort and safety of the patient [43], nurses must master many nursing skills in their careers. Patient transfer is one of the most significant nursing skills that they need to master. The process of patient transfer involves helping patients with mobility problems to move from a bed to a wheelchair. It is one of the most frequent and difficult nursing skills. According to a previous study, patient transfer is performed very frequently (on average, 26 times per nurse during a 4-h shift) in a hospital and nursing house [44]. During each step of patient transfer, nurses must apply proper body mechanics at appropriate times and also help the patient with the same. At the same time, nurses need to bear most of the weight of the patient. Incorrect procedures or postures may make the patient uncomfortable [45] and even cause injuries to both the patient and the nurse. Many nurses have been reported to suffer from occupational diseases, such as lower back pain, owing to the inadequacy of patient-transfer techniques [46], [47]. Therefore, we chose patient transfer as our target nursing skill to build the data set.

B. Fine-Gained Action Pairs

In this data set, we followed the setting of the former study [16]. We focused on the situation of basic patient transfer training for the beginner, i.e., transferring a patient from a bed to a wheelchair. According to the suggestions of experienced nursing teachers, the nurses were assumed to have the knowledge of basic patient transfer learned from textbooks and demonstration videos, while the patients were assumed to be weak person who cannot stand up by themselves. Based on these assumptions, basic patient transfer mainly involves three consecutive core action steps: 1) assisting the patient to stand up; 2) turn around; and 3) sit down (see Fig. 2). Also, the pattern of each nursing action was confined to common nursing training guideline, such as assuming that the nursing learner grips the lower back of the patient when assisting patient to stand up. We collected data samples of patient transfer considering a correct method and a typical incorrect method for performing each action step. It is worth noting that the incorrect method for each action step corresponds to the most common mistake made by the nursing learners according to the experience of professional nursing teachers.

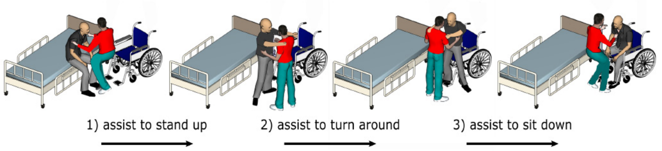


Fig. 2. Target nursing actions of patient transfer.

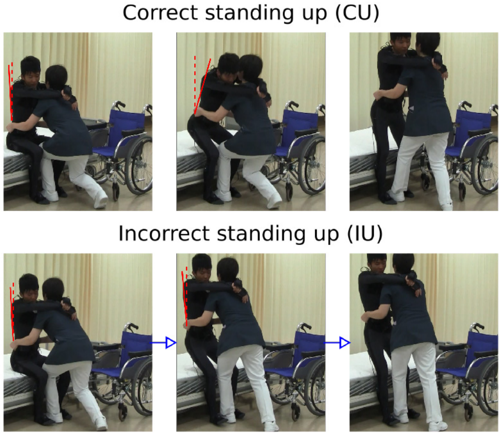
First, the correct and incorrect ways of assisting the patient in standing up are illustrated in Fig. 3. As for the correct way, the nurse needs to grab the clothes and bend the waist of the patient before helping them to stand up; the most common mistake made when assisting a patient to stand up is to allow them to stand up without bending their waist. The incorrect way makes it difficult for the nurse to complete the action and increases the risk of injuries and accidents. The correct and incorrect ways of assisting patients to turn around are illustrated in Fig. 4. The correct way for the nurse is to use the right foot as the turning pivot when the wheelchair is on the right side; the most common mistake made is to use the left foot as the turning pivot. Using the wrong foot as the pivot will cause inconvenience when turning around. Finally, the correct and incorrect ways of assisting a patient to sit down on a wheelchair are illustrated in Fig. 5. The most common mistake made when assisting the patient to sit down is to allow them to directly sit down on the wheelchair without bending their waist first. In total, there are six action step classes, denoted as cu, iu, ct, it, cd, and id, in our data set. For data set collection, nurses were asked to perform patient transfer correctly and incorrectly for each action step. During the patient transfer, a BSN with 17 IMU sensors was attached to the patient for kinematic information collection, including acceleration and rotational speed signals. The detailed configuration of our data set is presented in Section V-A.

IV. Proposed Method

First, we present an overview of the proposed method in Section IV-A. Then, we further introduce the multistage TCN, channel attention-based multistream structure, and label selector in Sections IV-B, IV-C, and IV-D, respectively. Finally, we present the loss functions in Section IV-E.

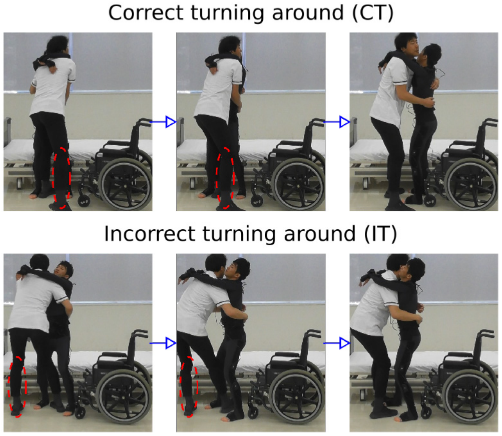
A. Overview

An overview of our proposed method is illustrated in Fig. 6. The importance of signal features from different forms of


Correct standing up (CU)

Incorrect standing up (IU)

Fig. 3. Correct/incorrect ways to assist the patient to stand up. When assisting the patient to stand in the incorrect way, the nurse pulls him or her directly upward without first bending the waist of the patient.

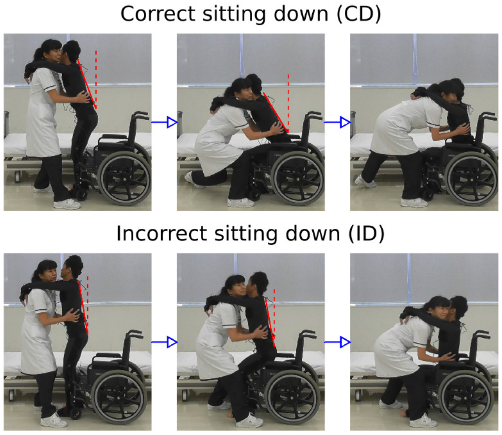

Correct turning around (CT)

Incorrect turning around (IT)

Fig. 4. Correct/incorrect ways to assist the patient to turn around. When assisting the patient to turn to the wheelchair in the incorrect way, the nurse uses the left foot as the turning pivot, which cannot be fixed during turning.

modalities will change over time. Different modalities, such as acceleration, rotational speed, and mixed modalities, have their own advantages for action detection of certain individual and certain action classes. It is helpful to focus more on the features from a particular form of modality that has more useful information for action recognition. Therefore, instead of using only one kind of modality by implementing the existing multistage TCN [25], we explore the potential of jointly using multiple forms of modalities and selecting the most reliable information from them. Based on this idea and the concept of the attention mechanism, we design a channel attention-based multistream architecture according to the existing multistage TCN.

In the first phase of the proposed architecture, we treat the time-series input of a form of modality as a stream branch. Specifically, we denote the time-series input signals from the IMU, including both acceleration and rotational speed signals, by $X_m = \{x_{m_1}, x_{m_2}, \ldots, x_{m_T}\}$, where T denotes the total time frames of all the time-series signals. $X_a = \{x_{a_1}, x_{a_2}, \ldots, x_{a_T}\}$ and $X_r = \{x_{r_1}, x_{r_2}, \ldots, x_{r_T}\}$ denote the acceleration and rotational speed signals, respectively. We consider X_r, X_a, and X_m as three wearable sensor input streams to three independent multistage TCN modules to give full play to the advantages of the respective modalities and their combination. Consequently, we can obtain framewise predictions $Y_r = \{y_{r_1}, y_{r_2}, \ldots, y_{r_T}\}$, $Y_a = \{y_{a_1}, y_{a_2}, \ldots, y_{a_T}\}$, and $Y_m = \{y_{m_1}, y_{m_2}, \ldots, y_{m_T}\}$ according to each input stream. Through the channel concatenation operation, we can obtain intermediate prediction features as $Y_f = \text{CAT}(Y_r, Y_a, Y_m)$. By applying channel-attention mechanisms to Y_f, we obtain the prediction features Y_c with the optimized channel weight, which means that the importance of prediction features from more convincing channels is magnified. Subsequently, we feed Y_c to one more multistage TCN module to generate the framewise prediction $Y_{fw} = \{y_1, y_2, \ldots, y_T\}$, which implies that an action class label (e.g., cu) will exist for each time frame. Finally, by using our label selector to filter the framewise predictions Y_{fw}, the action stepwise predictions $Y_{sw} = \{y_1, y_2, \ldots, y_S\}$ are generated, where $S = 3$ for our patient-transfer application.


Correct sitting down (CD)

Incorrect sitting down (ID)

Fig. 5. Correct/incorrect ways to assist the patient to sit down. When assisting the patient to sit in the incorrect way, the nurse allows the patient to fall vertically onto the wheelchair without bending the waist of the patient.

B. Multistage Temporal Convolutional Network

Not long ago, RNNs were the best choice for sequence modeling tasks because of their great ability to extract long-term temporal dependencies from time-series signals, such as LSTM [37] and GRU [48]. However, RNN-based recurrent models are empirically limited and difficult to train for actions defined by changes in features over too many time frames [24]. TCNs [39] were proposed to solve this problem by capturing longer and optimized temporal dependencies through stacked temporal (dilated) convolutional layers. Thus, we adopt a multistage TCN structure from [25] as our submodule to extract high-level temporal patterns from time-series data and features.

The structure of the multistage TCN is illustrated in Fig. 7. It consists of N stacked single stages. Each stage starts with a 1×1 convolutional layer to adjust the channel numbers of the input, followed by several stacked temporal convolutional layers with the same structure as that in [25]. The temporal convolutional layer mainly includes a 1-D dilated convolutional layer [49] with a kernel size of 3, a ReLU activation layer [50], a normal 1×1 convolutional layer, a dropout

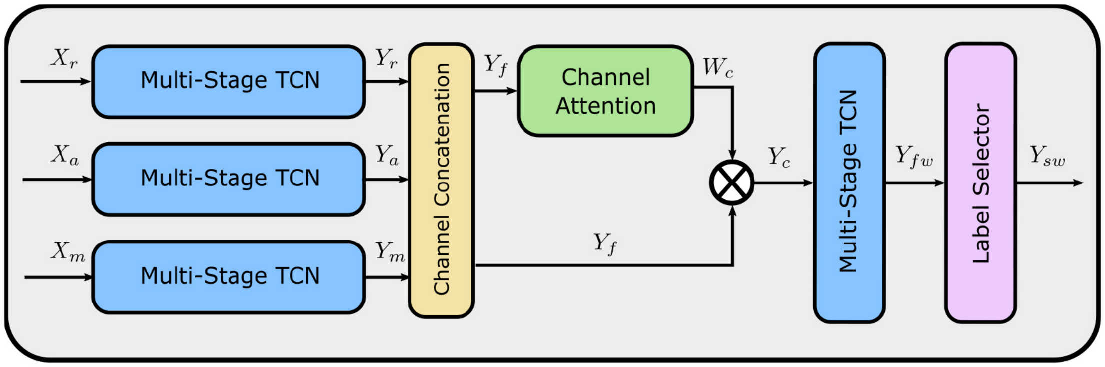


Fig. 6. *Architecture of the multistream* TCN. X_r, X_a, and X_m denote the different modal streams corresponding to rotational speed only, acceleration only, and mixed modalities, respectively; Y_r, Y_a, and Y_m denote the prediction features generated by independent multistage TCN modules (see Fig. 7); Y_f denotes the concatenation of multistream prediction features, Y_c denotes the prediction feature obtained by the channel-attention module (see Fig. 8), and Y_{fw} and Y_{sw} denote the framewise predictions and stepwise predictions generated by the label selector module (see Fig. 9).

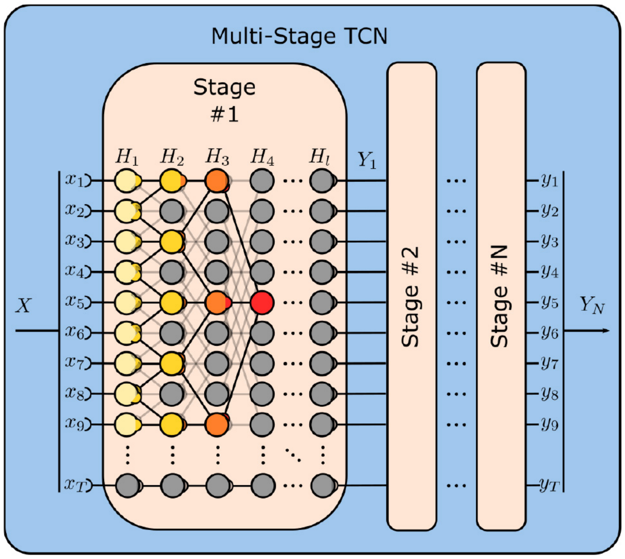

Multi-Stage TCN
Stage #1
H_1 H_2 H_3 H_4 H_l
Y_1
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_T
X
Stage #2
Stage #N
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8 y_9 y_T
Y_N

Fig. 7. Multistage TCN module. X represents the input of the multistage TCN module, which can be X_m, X_a, X_r, and Y_c. Y_n is the output of each stage. Y_N can be Y_m, Y_a, Y_r, and Y_fw. Each circle in each stage represents the feature map at each time frame. Each row of circles represents the output of a temporal convolutional layer as H_l.

layer [51], and a skip connection [52] from the input (i.e., the output from the last temporal convolutional layer). Thus, the process of one TCN can be expressed as follows:

$$r = 2^{l-1} \tag{1}$$
$$H_l^R = \text{ReLU}(W_1 *_r H_{l-1} + b_1) \tag{2}$$
$$H_l^D = \text{Dropout}\big(W_2 * H_l^R + b_2\big) \tag{3}$$
$$H_l = H_{l-1} + H_l^D \tag{4}$$

where r denotes the dilated rate of the dilated convolutional layer in the lth temporal convolutional layer; H_l denotes the output of the lth temporal convolutional layer; *Dropout* denotes the dropout operation; $*$ and $*_r$ denote the standard convolutional operation and dilated convolutional operation with dilated rate r, respectively; W_1 and W_2 denote the weights of the dilated and standard convolutional layers, respectively; and b_1 and b_2 denote bias vectors. According to the above equations, the dilated factor is doubled with increase in the number of temporal convolutional layers $(1, 2, 4, \ldots, 2^l)$. Thus, as shown in Fig. 7, the farther the position of the temporal convolutional layer, the larger its temporal receptive field. In other words, a farther layer can access a larger range of optimized temporal dependencies from time-series signals. Finally, the output of the last temporal convolutional layer is the output of that stage, given by

$$H_{n,L} = \mathcal{S}_n\big(H_{n,0}\big) \tag{5}$$

where L denotes the total number of temporal convolutional layers in the stage; $H_{n,L}$ and $H_{n,0}$ denote the output of the last temporal convolutional layer in the nth stage and the input of this stage, respectively; and $\mathcal{S}_n$ denotes the function of the nth single temporal convolutional stage.

From the perspective of the stage, we can obtain the intermediate prediction Y_n made by adding the SoftMax function to the output of the stage $H_{n,L}$. Then, the intermediate prediction of the current stage will be refined by feeding itself as input to the next stage. Thus, the process of feeding the input to N stacked stages is described as follows:

$$\begin{aligned} Y_0 &= X && (6)\\ Y_n &= \text{SoftMax}(\mathcal{S}_n(Y_{n-1})),\ \ n \in \{1, \ldots, N\}\\ &= \text{SoftMax}\big(H_{n,L}\big),\ \ n \in \{1, \ldots, N\}. && (7)\end{aligned}$$

Finally, the output Y_N of the last stage is the final output of that multistage TCN module

$$Y_N = \text{MSTCN}(X) \tag{8}$$

where MSTCN$(\cdot)$ denotes the function of the multistage TCN, and X denotes the input.

In our architecture, there are three input streams: X_r, X_a, and X_m. These represent an input stream with only rotational speed signals, an input stream with only acceleration signals, and an input stream with mixed signals of the former two streams, respectively. Through individual multistage TCN modules, we can obtain prepredictions for each stream as $Y_r =$ MSTCN(X_r), Y_a = MSTCN(X_a), and Y_m = MSTCN(X_m), respectively.

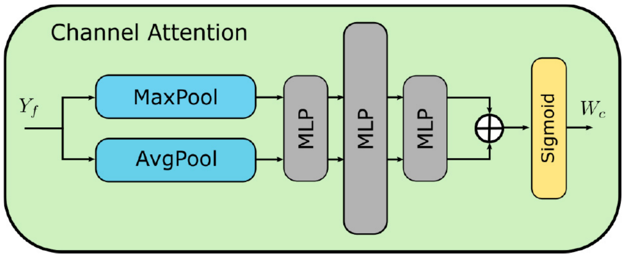


Fig. 8. Channel-attention module. Y_f denotes the concatenated intermediate prediction features, and W_c denotes the channel weight vector.

C. Channel Attention-Based Multistream Structure

To integrate the advantages of prediction features from different input streams, we adopt the channel-attention mechanism [42] for attaching an attention weight to each channel of the intermediate prediction features Y_f. The structure of our channel-attention module is illustrated in Fig. 8. Here, average pooling and max pooling are used together to compute temporal statistics and distinctive object features by downsampling the temporal dimension of input features as one with an average filter and a max filter, respectively.

By feeding the intermediate feature F_y to these pooling layers, we can obtain the features M_{avg} and M_{max}, respectively. Then, the channel-attention map W_c can be generated by a weight-shared network. The outputs of forwarding M_{avg} and M_{max} to a multilayer perceptron (MLP) will be merged by elementwise summation. Finally, the sigmoid function is applied to constrain the attention-weight values in a relatively small range. The entire process of the channel-attention module can be expressed as follows:

$$M_{\text{max}} = \text{MaxPool}(Y_f) \tag{9}$$

$$M_{\text{avg}} = \text{AvgPool}(Y_f) \tag{10}$$

$$W_c = \sigma(WM_{\text{max}} + WM_{\text{avg}} + b) \tag{11}$$

where σ denotes the sigmoid function, and W and b denote the weights and bias vectors of the MLP, respectively.

D. Label Selector

The framewise prediction results may not be intuitive or user friendly for nurses, especially when oversegmentation occurs. Considering the application purpose, we propose a label selector to extract valid action class labels from framewise predictions, named stepwise predictions. Here, we assume that the number of performed action steps is fixed at three. The label selector counts the number of framewise labels in each action step class as follows:

$$N_c = \sum_{t=1}^{T} \text{Equal}(\text{ArgMax}(y_t), c) \tag{12}$$

$$\text{Equal}(y_1, y_2) = \begin{cases} 1, & \text{if } y_1 = y_2 \\ 0, & \text{else} \end{cases} \tag{13}$$

where $c \in C$ with $C = \{cu, iu, ct, it, cd, id\}$, $y_t \in Y_{fw}$, denotes the prediction for each time frame, and $\text{ArgMax}(y_t)$ denotes the class label with the largest probability value at time frame t. The label selector saves three action labels with the largest counts. Furthermore, it calculates the "center of mass" for each of these labels as follows:

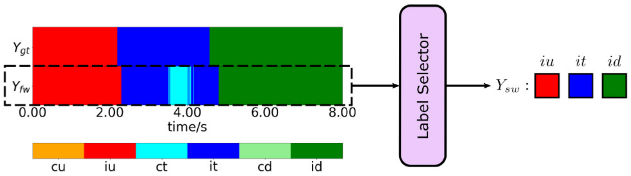


Fig. 9. Label selector module.

$$\xi_c = \frac{\sum_{n=1}^{N_c} t_n}{N_c} \tag{14}$$

where t_n denotes the time frame for each prediction that belongs to class c. According to ξ_c, the label selector can determine the order of selected labels from small to large. Consider Fig. 9 as an example. The proposed label selector can extract a stepwise prediction of iu-it-id (red-blue-green) from the framewise prediction.

E. Loss Function

For the loss function, a combination of classification loss and smoothing loss is applied to the output of each temporal convolutional stage as our optimization goal

$$\mathcal{L} = \sum_s (\mathcal{L}_{\text{cls,s}} + \lambda \mathcal{L}_{\text{smt,s}}) \tag{15}$$

where λ denotes the hyperparameter that adjusts the contributions of different loss functions. The classification loss is defined as the cross-entropy loss as follows:

$$\mathcal{L}_{\text{cls}} = \frac{1}{T} \sum_{t=1}^{T} -\log(\widehat{y}_{t,gt}) \tag{16}$$

where $\widehat{y}_{t,gt}$ denotes the predicted probability for the ground-truth class at time frame t. Additionally, we adopt a smoothing loss [25] to reduce the over-segmentation errors as follows:

$$\mathcal{L}_{\text{smt}} = \frac{1}{TC} \sum_{t,c} \Delta_{t,c}^2 \tag{17}$$

$$\Delta_{t,c} = \begin{cases} \left|\log \frac{y_{t,c}}{y_{t-1,c}}\right|, & \Delta_{t,c} \le \tau \\ \tau, & \text{else} \end{cases} \tag{18}$$

where $y_{t,c}$ denotes the predicted probability for class c at time frame t, and τ is the smoothing loss threshold.

V. Experiment and Results

First, we present experimental implementation details in Section V-A. The comparison results and performance on nursing skill data set are detailed in Section V-B. Furthermore, the effectiveness of the proposed channel attention-based multistream structure is evaluated in Section V-C. The effect of using different combinations of sensor placements is explored in Section V-D. Finally, we also show comparison results on publicly available data set, C-MHAD.

TABLE I
DETAILS OF THE NURSING SKILL ASSESSMENT DATA SET

Nurse	cu-ct-cd	cu-it-cd	cu-ct-id	cu-it-id	iu-ct-cd	iu-it-cd	iu-ct-id	iu-it-id	Total
#1	4	4	4	4	4	4	4	4	32
#2	4	4	4	4	4	4	4	4	32
#3	4	4	4	4	4	4	4	4	32
#4	4	4	4	4	4	4	4	4	32
#5	4	4	4	4	4	4	4	4	32
#6	4	4	4	4	4	4	4	4	32
Total	24	24	24	24	24	24	24	24	192

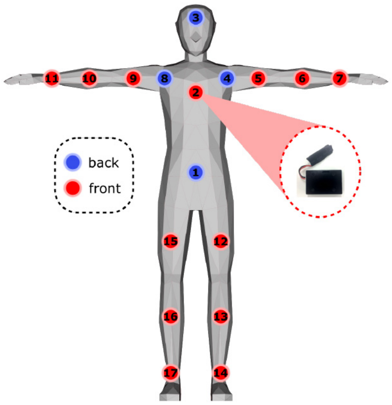


Fig. 10. Placements of IMU sensors on the body of the patient. In total, 17 IMU sensors were integrated into our BSN. The blue and red nodes represent the placements at the back and front, respectively.

A. Implementation Details

Our patient-transfer data set contains a total of eight combinations of consecutive action steps (e.g., iu-ct-cd). Overall, six nurses (four females and two males) were asked to perform four trials of each possible combination on the same male patient, as summarized in Table I. Based on previous works [53]–[56] on the effect of sensor placement for HAR tasks, different actions prefer different sensor placements. For example, the waist and thighs are considered as the best placements for fall detection [57], [58]. In our case, we used a total of 17 IMU sensors (ZMP, IMU-Z2) to cover the optimal placements mentioned in the previous studies, such as waist, wrist, chest, arm, thigh, ankle, and so on. These sensors were attached to the clothing worn by the patient. Inevitably, these sensors will move slightly during patient transfer. However, data-driven methods trained on the data from these sensors should learn how to robustly handle these movement disturbances. The placement of each IMU sensor and the corresponding sensor number is shown in Fig. 10. Each IMU has six kinematic variables (channels), including acceleration and rotational speed, in the x, y, and z axes. The working frequency was 50 Hz. Thus, there is a 102-dimensional vector that contains the kinematic information at each time frame. Finally, we obtained 24 samples for each nurse and 32 samples for each combination of consecutive actions. Overall, there were 192 samples for patient transfer and 576 samples for all action steps.

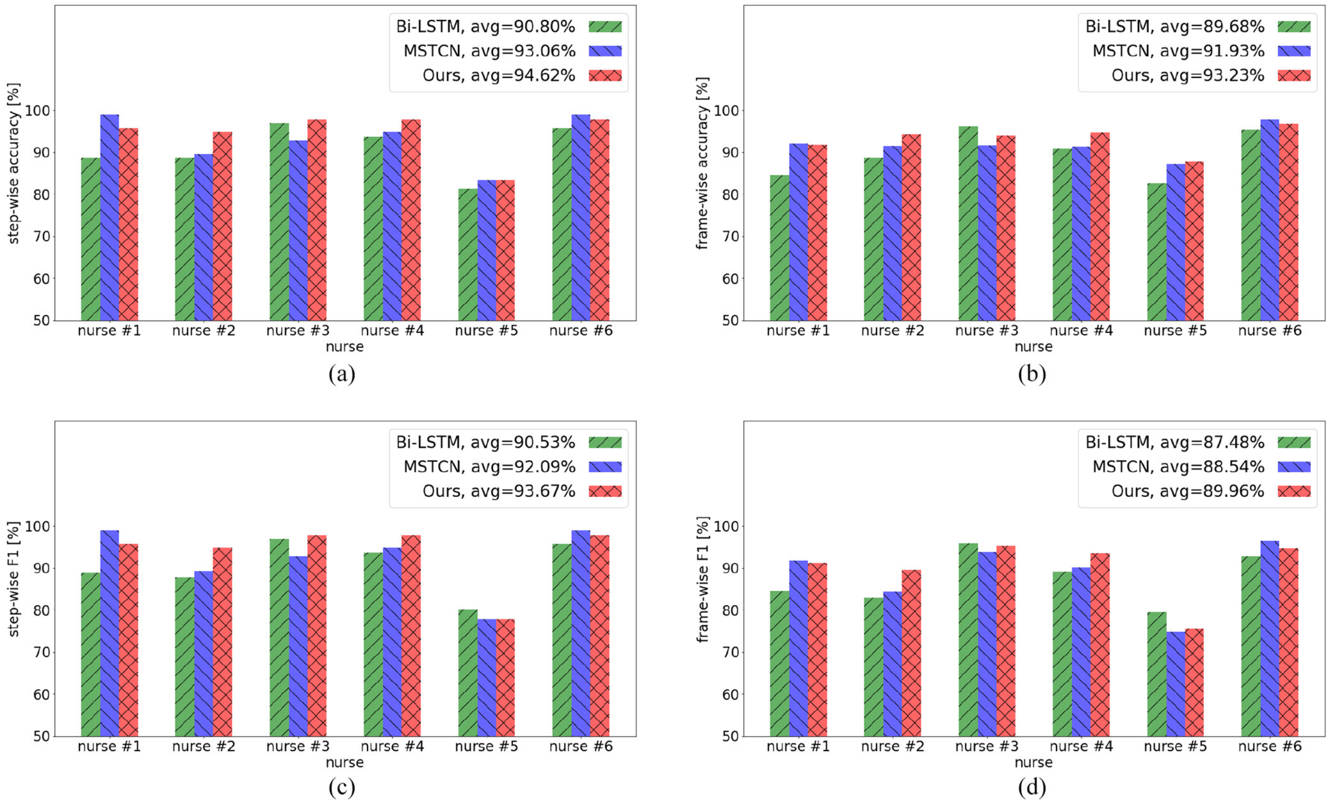
TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON NURSING THE SKILL ASSESSMENT DATA SET. BOLD FONT INDICATES THE BEST PERFORMANCE

	Acc_{step}	$F1_{step}$	Acc_{frame}	$F1_{frame}$
DCNN [17] (SW)	50.7%	41.8%	-	-
HMM [33] (SW)	76.6%	74.3%	-	-
Catal *et al.* [32] (SW)	78.1%	77.2%	-	-
DRNN [20] (SW)	79.2%	77.9%	-	-
MA-RNN [21] (SW)	82.6%	81.7%	-	-
Bi-LSTM [23]	90.8%	90.5%	89.7%	87.5%
MSTCN [25]	93.1%	92.1%	91.9%	88.5%
Ours	**94.6%**	**93.7%**	**93.2%**	**90.0%**

Leave-one-subject-out cross-validation (LOSOXV) was used to evaluate the performance of the proposed method and other state-of-the-art action detection methods, including the SW-based and seq2seq schemes. We implemented an SW-based scheme using DCNN [17], HMM [33], the ensemble model of the J48 decision tree, MLP and logistic regression [32], DRNN [20], and MA-RNN [21]. Each of these models was implemented as described in the corresponding paper. The SW size was fine-tuned for each model based on our data set for better performance. For HMM, the window size was 0.4 s, whereas the window size of the remaining models was 0.8 s. All the overlapping sizes are half of the window size. As for the seq2seq scheme, to the best of our knowledge, all the existing models, including Bi-LSTM [23] and MSTCN [25], are designed for video signals. Guided by the paper and original code , we attempted to fine-tune the hyperparameters and structures of these two methods [23], [25] to better handle wearable sensor signals. The main hyperparameters of our method were set empirically. We set the numbers of stages of multistage TCN before and after channel attention at 3 and 1, respectively, with ten temporal convolutional layers in each stage. The number of output channels for each convolutional layer in our model was 256. The dropout rate was set at 0.3. With regard to the loss functions, λ and τ were set at 0.15 and 4, respectively. We trained the method using the ADAM optimizer [59] with a learning rate of 10^{-3}. The number of epochs and batch size was 30 and 32, respectively. Furthermore, the stepwise accuracy ($\mathrm{Acc}_{\mathrm{step}}$) and F1 score ($F1_{\mathrm{step}}$), as well as the framewise accuracy ($\mathrm{Acc}_{\mathrm{frame}}$) and F1 score ($F1_{\mathrm{frame}}$), were used as the metrics in this study.

B. Comparison on Nursing Skill Assessment Data Set

The results of the comparison between the proposed method and the state-of-the-art SW-based and seq2seq action detection methods on our nursing skill assessment data set are presented in Table II). In the case of SW-based methods, the framewise accuracy and F1 score are easily affected by the window size, and the prediction areas overlap because the step size of the SW is smaller than the window size. Thus, we only generated



(a) (b) (c) (d)

Fig. 11. Comparison of accuracy and F1 score between our method and state-of-the-art methods for each nurse. (a) Stepwise accuracy. (b) Framewise accuracy. (c) Stepwise F1 score. (d) Framewise F1 score.

the stepwise accuracy and F1 score for the SW-based methods. As for the seq2seq methods, both framewise and stepwise performances are detailed. We can observe that seq2seq methods significantly outperformed SW-based methods in terms of all the metrics. The experimental results demonstrate that the seq2seq methods are superior in terms of action detection, which is a result of their optimal temporal receptive field. Our method outperforms the seq2seq methods, achieving a stepwise accuracy of 94.62%, stepwise F1 score of 93.67%, framewise accuracy of 93.23%, and framewise F1 score of 89.96%.

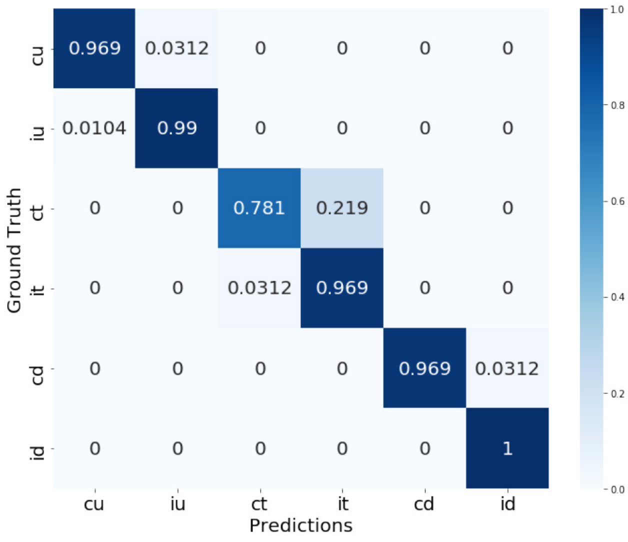


Fig. 12. Stepwise confusion matrix of our method on the nursing skill assessment data set.

We also illustrate the performance of each seq2seq method for each nurse in Fig. 11. The performances of the methods vary between individuals. All the methods deliver the worst performance in the case of nurse #5. Nevertheless, our method clearly outperforms the other methods in the cases of more individuals. The average confusion matrix for different nurses using our method is presented in Fig. 12. It demonstrates that our method only makes mispredictions between the fine-grained action class pairs, i.e., the correct and incorrect ways of performing the same nursing action step. This indicates that reducing the mispredictions between fine-grained action class pairs is the key to improve the accuracy. The correctness of assisting a patient to turn around (ct and it), which accounts for 77.46% of all the mispredictions, is more difficult to assess than the correctness of the other two action steps. However, from the confusion matrix for each individual, we find that all *ct* actions for nurse #5 are wrongly predicted as *it*, which is the main reason for the low accuracy of correctness assessment of assisting a patient to turn around. This demonstrates the importance of appropriately addressing individual differences in action detection tasks. Data samples of individuals such as nurse #5, whose pattern is fairly different from those of other nurses, are unavoidable. The recorded video suggests that nurse #5 is considerably gentler than other nurses when

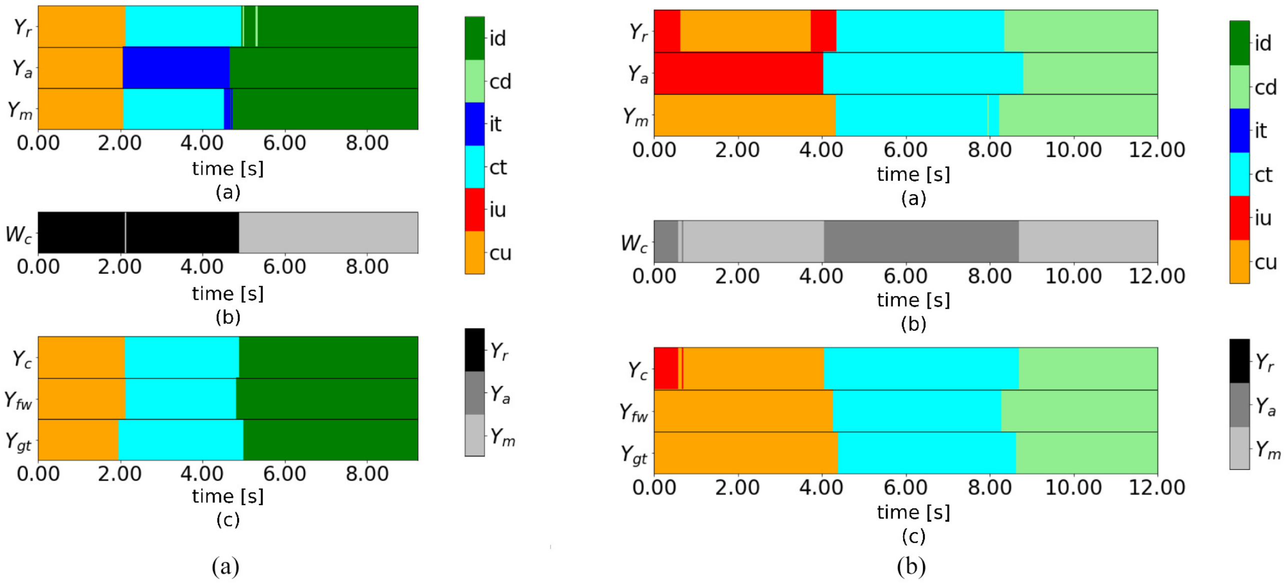


Fig. 13. Visualization of channel attention for the multistream structure. Rows Y_r, Y_a, and Y_m represent framewise predictions according to each modal stream; row W_c represents the most convincing stream at each time frame; Y_c represents framewise predictions considering W_c; and Y_{fw} and Y_{gt} represent final framewise predictions of our method and the corresponding ground truth, respectively. (a) Example #1. (b) Example #2.

TABLE III

FURTHER COMPARISON WITH MSTCN USING DIFFERENT STREAMS AS INPUT. BOLD FONT INDICATES THE BEST PERFORMANCE

	Acc_{step}		$F1_{step}$		Acc_{frame}		$F1_{frame}$	
	avg.	std.	avg.	std.	avg	std.	avg.	std.
a-MSTCN	93.1%	6.6×10^{-2}	92.0%	8.4×10^{-2}	91.8%	5.8×10^{-2}	87.0%	9.1×10^{-2}
r-MSTCN	92.9%	5.7×10^{-2}	91.9%	7.6×10^{-2}	91.7%	3.6×10^{-2}	88.6%	7.5×10^{-2}
MSTCN	93.1%	5.5×10^{-2}	92.1%	7.2×10^{-2}	91.9%	3.1×10^{-2}	88.5%	7.2×10^{-2}
Ours	**94.6%**	$\mathbf{5.2 \times 10^{-2}}$	**93.7%**	$\mathbf{7.2 \times 10^{-2}}$	**93.2%**	$\mathbf{2.8 \times 10^{-2}}$	**90.0%**	$\mathbf{6.8 \times 10^{-2}}$

performing actions classified as incorrect. Furthermore, the use of the wrong foot as the pivot when assisting the patient to turn around mainly affects the state of the nurse rather than that of the patient, but BSNs are only placed on the patient. These may be the reasons why none of the methods performed well for this action of nurse #5.

C. Effectiveness of Multistream Structure

The proposed method is novel in that a channel attention-based multistream structure for multistage TCN modules was designed to effectively utilize the multimodal signals of the IMU sensor. Thus, to demonstrate the effectiveness of our multistream structure, we further compared our method with the pure multistage TCN (i.e., MSTCN) by using different signal modalities. In addition to using full signals, we tested the signals with respect to only rotational speed and only acceleration on MSTCN , denoted by r-MSTCN and a-MSTCN, respectively. We can directly observe that the use of full data with mixed modalities only achieves a performance similar to those of r-MSTCN and a-MSTCN (see Table III). However, our proposed method can outperform the pure MSTCN because it dynamically considers the characteristics of each modality of the IMU signals. Our method obtains the lowest values of the standard deviations of both framewise and stepwise accuracies and F1 scores for all the nurses, as presented in Table III. This implies that our method is more stable than pure MSTCN in the presence of individual differences.

To further demonstrate the effectiveness of the proposed structure, we visualized the output of each critical module in our method (see Fig. 13). Rows Y_r, Y_a, and Y_m in both examples #1 [see Fig. 13(a)] and #2 [see Fig. 13(b)] represent the labels with the largest probability at each time frame; row W_c represents the stream to which the largest predicted probability belongs to, or in other words, the most convincing stream at that time frame; row Y_c represents the label with the largest probability from the output of the most convincing stream; and rows Y_{fw} and Y_{gt} represent the framewise predictions of our method and the corresponding framewise ground truth, respectively. First, for example, #1, the oversegmentation of Y_r occurs between the steps of turning around and sitting down. All the predictions of Y_a and part of the predictions of Y_m are incorrect for the turning step, shown in blue (it) instead of light blue (ct). According to row W_c, our method mainly focuses on the predictions of rotational speed stream Y_r for the first half of the data signals and those of the mixed modalities stream Y_m for the second half of the data signals. Thus, Y_c can avoid the oversegmentation and misprediction problems of

TABLE IV
PERFORMANCE OF OUR METHOD USING DIFFERENT SENSOR COMBINATIONS

Combination	Acc_{step}	$F1_{step}$	Acc_{frame}	$F1_{frame}$
1-17	94.6%	93.7%	93.2%	90.0%
1, 2, 4, 8, 12, 15	92.4%	91.6%	92.0%	88.0%
1, 7, 11, 12, 15	87.7%	87.2%	86.0%	85.7%
7, 11, 12, 15	79.0%	78.4%	79.4%	79.7%
1, 7, 11	86.8%	85.4%	83.3%	83.5%
1, 12, 15	85.2%	84.1%	86.7%	81.3%
7, 11	66.3%	60.9%	63.1%	60.6%
12, 15	67.2%	63.3%	73.6%	65.6%
1	76.9%	74.0%	78.5%	71.9%

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS ON C-MHAD, IN TERMS OF F1 SCORE. BOLD FONT INDICATES THE BEST PERFORMANCE

	Transition Movements	TV Gestures	Data
Wei *et al.* [60]	56.4%	60.3%	inertial
Wei *et al.* [60]	75.7%	77.8%	video
Wei *et al.* [60]	78.8%	81.8%	inertial&video
HMM [33]	28.0%	57.4%	inertial
Catal *et al.* [32]	50.0%	62.8%	inertial
DRNN [20]	68.8%	81.1%	inertial
MA-RNN [21]	82.0%	78.7%	inertial
Bi-LSTM [23]	89.7%	96.1%	inertial
MSTCN [25]	93.4%	97.0%	inertial
Ours	**95.3%**	**98.5%**	inertial

each stream, and help generate the framewise prediction Y_{fw}. In example #2, we can directly observe that the prediction features of Y_m are the most useful for making predictions with this sample. Our method mainly uses the prediction features of the acceleration stream Y_a and mixed modalities stream Y_m for this example. Although mispredictions, shown in red (iu), still exist at the beginning of Y_c, the latter multistage TCN module helps further restore these parts, shown in orange (cu).

Both quantitative and visualization results reveal that the proposed method can better leverage the advantages of different modalities of IMU signals. It achieves a higher recognition accuracy than those of state-of-the-art methods.

D. Sensor Combination Exploration

The number of sensors is typically constrained based on the application. Additionally, sensor placements are not arbitrary because real-world applications need to consider the feasibility of sensor placements. Hence, it is important to explore the effects of using different placement combinations of IMU sensors, illustrated in Fig. 10, especially for some meaningful placements. The results of using different wearable sensor combinations are listed in Table IV. First, if we only use the sensors at the trunk of the body, which are numbered 1, 2, 4, 8, 12, and 15, the performance of our model is comparable to that achieved when using the data of all 17 sensors. In addition, we choose some meaningful placements of IMU sensors that are likely to be adopted in daily life. Examples include placements 7 and 11, where one may wear a smartwatch, and placements 12, 15, and 1, where one may place their smartphone. The results reveal that the placements at relatively passive parts of the body of the patient, such as 1, are better choices for assessing nursing skills, such as patient transfer, when the number of IMUs is constrained. Furthermore, the results indicate that our method has the potential to assess nursing or other skills using only everyday devices with embedded IMU sensors (e.g., smartphones).

E. Comparison on Public Data Set (C-MHAD)

To show effectiveness and robustness of the proposed method, we also did experiments on publicly available data set, C-MHAD [60] (continuous multimodal human action dataset). C-MHAD consists of two action sets: 1) transition movements (e.g., stand-to-sit, stand-to-fall, lie-to-sit, etc.), seven action classes in total and 2) smart TV gestures (e.g., draw-clockwise-circle, swipe-to-left, right-hand wave, etc.), five action classes in total. In C-MHAD, 12 subjects performed ten trials of randomized continuous actions (2 min per trial) within the action set.

We followed the training and test data set setting and measurement metrics of the benchmark performance [60] to conduct experiments. In addition to the benchmark performance provided by the data set authors, including the use of inertial data, video data, and their fusion, our and other methods are based on inertial data only. The results are shown in the Table V, in terms of F1 score. Regarding transition movements of C-MHAD, we correctly detected and recognized 102 out of a total of 107 test actions from 12 subjects. Most of the mis-predictions occurred between similar action classes, i.e., fine-grained action pairs, such as lie-to-stand and lie-to-sit, sit-to-lie and stand-to-lie, etc. Regarding smart TV gestures of C-MHAD, we correctly detect and recognize 99 out of a total of 100 test actions from 12 subjects. Mis-predictions were mainly caused by interference from irrelevant gestures. Our method achieves the best performance, using only inertial data, with F1 score of 95.3% for the action set of transition movements and 98.5% for the action set of smart TV gestures. Our results even surpass the benchmark performance using both inertial and video fusion data (F1 score of 78.8% for transition movements and 81.8% for smart TV gestures) by a large margin.

VI. CONCLUSION AND FUTURE WORKS

Herein, we proposed an innovative seq2seq method for wearable sensor-based fine-grained action detection. In contrast to the existing seq2seq action detection methods, the proposed method can utilize the advantages of each modal signal or a combination of modal signals in the case of distinct actions of different individuals. We realized the proposed method by incorporating a multistage TCN module into a channel attention-based multistream structure. Each stream in the structure represents a form of modality or a combination of modalities. We applied these BSN-based fine-grained detection techniques for a promising application, namely, automatic nursing skill assessment. By collecting a nursing skill data set of patient transfer using BSNs, the proposed method realized automatic nursing skill assessment with average stepwise and framewise accuracies of 94.62% and 93.23%, respectively. The proposed method outperformed the state-of-the-art action detection methods on our nursing skill data set and public data set (C-MHAD) owing to its ability to leverage the features of more convincing IMU modalities.

In the future, we plan to collect a larger data set with more nursing actions and skills. We will further consider how to provide nursing learners with more informative feedback in addition to correction information. Furthermore, lack of training data may be an inevitable problem encountered by supervised machine learning applications such as ours. The quantity of labeled data samples is small owing to the cost of data acquisition and privacy concerns. We will aim to determine how a reliable nursing skill assessment system can be built based on a small data set.

REFERENCES

[1] M. Kranz *et al.*, “The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices,” *Pervasive Mobile Comput.*, vol. 9, no. 2, pp. 203–215, Apr. 2013.

[2] D. Yang *et al.*, “TennisMaster: An IMU-based online serve performance evaluation system,” in *Proc. 8th Augmented Human Int. Conf.*, 2017, pp. 1–8.

[3] M. J. Fard, S. Ameri, R. D. Ellis, R. B. Chinnam, A. K. Pandya, and M. D. Klein, “Automated robot-assisted surgical skill evaluation: Predictive analytics approach,” *Int. J. Med. Robot. Comput. Assist. Surgery*, vol. 14, no. 1, p. e1850, 2018.

[4] R. Hou, C. Chen, and M. Shah, “An end-to-end 3D convolutional neural network for action detection and segmentation in videos,” 2017. [Online]. Available: arXiv:1712.01111.

[5] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, “A database for fine grained activity detection of cooking activities,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1194–1201.

[6] G. Fortino, S. Galzarano, R. Gravina, and W. Li, “A framework for collaborative computing and multi-sensor data fusion in body sensor networks,” *Inf. Fusion*, vol. 22, pp. 50–70, Mar. 2015.

[7] S. Movassaghi, M. Abolhasan, J. Lipman, D. Smith, and A. Jamalipour, “Wireless body area networks: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1658–1686, 3rd Quart., 2014.

[8] S. Ullah *et al.*, “A comprehensive survey of wireless body area networks,” *J. Med. Syst.*, vol. 36, no. 3, pp. 1065–1094, 2012.

[9] C. Feng, S. Arshad, S. Zhou, D. Cao, and Y. Liu, “Wi-Multi: A three-phase system for multiple human activity recognition with commercial WiFi devices,” *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7293–7304, Aug. 2019.

[10] A. Ahmadi *et al.*, “Toward automatic activity classification and movement assessment during a sports training session,” *IEEE Internet Things J.*, vol. 2, no. 1, pp. 23–32, Feb. 2015.

[11] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, “Deep convolutional neural networks on multichannel time series for human activity recognition,” in *Proc. IJCAI*, vol. 15, 2015, pp. 3995–4001.

[12] M. Marć, A. Bartosiewicz, J. Burzyńska, Z. Chmiel, and P. Januszewicz, “A nursing shortage—A prospect of global and local policies,” *Int. Nursing Rev.*, vol. 66, no. 1, pp. 9–16, 2019.

[13] S. Joshi, “Coronavirus disease 2019 pandemic: Nursing challenges faced,” *Cancer Res. Stat. Treatment*, vol. 3, no. 5, pp. 136–137, 2020.

[14] D. A. Nardi and C. C. Gyurko, “The global nursing faculty shortage: Status and solutions for change,” *J. Nursing Scholarship*, vol. 45, no. 3, pp. 317–326, 2013.

[15] Z. Huang *et al.*, “Self-help training system for nursing students to learn patient transfer skills,” *IEEE Trans. Learn. Technol.*, vol. 7, no. 4, pp. 319–332, Oct.–Dec. 2014.

[16] Z. Huang *et al.*, “Automatic evaluation of trainee nurses’ patient transfer skills using multiple kinect sensors,” *IEICE Trans. Inf. Syst.*, vol. 97, no. 1, pp. 107–118, 2014.

[17] W. Jiang and Z. Yin, “Human activity recognition using wearable sensors by deep convolutional neural networks,” in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1307–1310.

[18] H. Zhang, Z. Xiao, J. Wang, F. Li, and E. Szczerbicki, “A novel IoT-perceptive human activity recognition (HAR) approach using multihead convolutional attention,” *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1072–1080, Feb. 2020.

[19] D. Tao, Y. Wen, and R. Hong, “Multicolumn bidirectional long short-term memory for mobile devices-based human activity recognition,” *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1124–1134, Dec. 2016.

[20] A. Murad and J.-Y. Pyun, “Deep recurrent neural networks for human activity recognition,” *Sensors*, vol. 17, no. 11, p. 2556, 2017.

[21] Z. Zhong, C. Lin, T. Ogata, and J. Ota, “Multi-attention deep recurrent neural network for nursing action evaluation using wearable sensor,” in *Proc. 25th Int. Conf. Intell. User Interfaces*, 2020, pp. 546–550.

[22] D. Chen *et al.*, “Bring gait lab to everyday life: Gait analysis in terms of activities of daily living,” *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1298–1312, Feb. 2020.

[23] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, “A multi-stream bi-directional recurrent neural network for fine-grained action detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1961–1970.

[24] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks for action segmentation and detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 156–165.

[25] Y. A. Farha and J. Gall, “MS-TCN: Multi-stage temporal convolutional network for action segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3575–3584.

[26] R. Poppe, “A survey on vision-based human action recognitionm,” *Image Vision Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[27] H.-B. Zhang *et al.*, “A comprehensive survey of vision-based human action recognition methods,” *Sensors*, vol. 19, no. 5, p. 1005, 2019.

[28] Y. Gu, F. Ren, and J. Li, “PAWS: Passive human activity recognition based on WiFi ambient signals,” *IEEE Internet Things J.*, vol. 3, no. 5, pp. 796–805, Oct. 2016.

[29] J. Yang, H. Zou, H. Jiang, and L. Xie, “Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes,” *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3991–4002, Oct. 2018.

[30] I. C. Gyllensten and A. G. Bonomi, “Identifying types of physical activity with a single accelerometer: Evaluating laboratory-trained algorithms in daily life,” *IEEE Trans. Biomed. Eng.*, vol. 58, no. 9, pp. 2656–2663, Sep. 2011.

[31] L. Fan, Z. Wang, and H. Wang, “Human activity recognition model based on decision tree,” in *Proc. Int. Conf. Adv. Cloud Big Data*, 2013, pp. 64–68.

[32] C. Catal, S. Tufekci, E. Pirmit, and G. Kocabag, “On the use of ensemble of classifiers for accelerometer-based activity recognition,” *Appl. Soft Comput.*, vol. 37, pp. 1018–1022, Dec. 2015.

[33] J. Wang, R. Chen, X. Sun, M. F. H. She, and Y. Wu, “Recognizing human daily activities from accelerometer signal,” *Procedia Eng.*, vol. 15, pp. 1780–1786, Jan. 2011.

[34] H. Gjoreski, J. Bizjak, M. Gjoreski, and M. Gams, “Comparing deep and classical machine learning methods for human activity recognition using wrist accelerometer,” in *Proc. IJCAI Workshop Deep Learn. Artif. Intell.*, vol. 10. New York, NY, USA, 2016, p. 970.

[35] F. J. Ordóñez and D. Roggen, “Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition,” *Sensors*, vol. 16, no. 1, p. 115, 2016.

[36] Y. Zhao, R. Yang, G. Chevalier, X. Xu, and Z. Zhang, “Deep residual bidir-lstm for human activity recognition using wearable sensors,” *Math. Problems Eng.*, vol. 2018, no. 9, 2018, Art. no. 7316954. [Online]. Available: https://doi.org/10.1155/2018/7316954

[37] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.

[39] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," 2016. [Online]. Available: arXiv:1609.03499.

[40] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014. [Online]. Available: arXiv:1409.0473.

[41] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[43] B. Schibye, A. F. Hansen, C. T. Hye-Knudsen, M. Essendrop, M. Böcher, and J. Skotte, "Biomechanical analysis of the effect of changing patient-handling technique," *Appl. Ergonom.*, vol. 34, no. 2, pp. 115–123, 2003.

[44] A. Garg, B. D. Owen, and B. Carlson, "An ergonomic evaluation of nursing assistants' job in a nursing home," *Ergonomics*, vol. 35, no. 9, pp. 979–995, 1992.

[45] K. Kjellberg, M. Lagerström, and M. Hagberg, "Patient safety and comfort during transfers in relation to nurses' work technique," *J. Adv. Nursing*, vol. 47, no. 3, pp. 251–259, 2004.

[46] A. Karahan and N. Bayraktar, "Determination of the usage of body mechanics in clinical settings and the occurrence of low back pain in nurses," *Int. J. Nursing Stud.*, vol. 41, no. 1, pp. 67–75, 2004.

[47] W. S. Marras, K. G. Davis, B. C. Kirking, and P. K. Bertsche, "A comprehensive analysis of low-back disorder risk and spinal loading during the transferring and repositioning of patients using different techniques," *Ergonomics*, vol. 42, no. 7, pp. 904–926, 1999.

[48] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014. [Online]. Available: arXiv:1406.1078.

[49] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015. [Online]. Available: arXiv:1511.07122.

[50] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Stat.*, 2011, pp. 315–323.

[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[53] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Proc. Int. Conf. Pervasive Comput.*, 2004, pp. 1–17.

[54] L. Atallah, B. Lo, R. King, and G.-Z. Yang, "Sensor positioning for activity recognition using wearable accelerometers," *IEEE Trans. Biomed. Circuits Syst.*, vol. 5, no. 4, pp. 320–329, Aug. 2011.

[55] I. Orha and S. Oniga, "Study regarding the optimal sensors placement on the body for human activity recognition," in *Proc. IEEE 20th Int. Symp. Design Technol. Electron. Packag. (SIITME)*, 2014, pp. 203–206.

[56] N. Jablonsky, S. McKenzie, S. Bangay, and T. Wilkin, "Evaluating sensor placement and modality for activity recognition in active games," in *Proc. Aust. Comput. Sci. Week Multiconf.*, 2017, pp. 1–8.

[57] P. Ntanasis, E. Pippa, A. T. Özdemir, B. Barshan, and V. Megalooikonomou, "Investigation of sensor placement for accurate fall detection," in *Proc. Int. Conf. Wireless Mobile Commun. Healthcare*, 2016, pp. 225–232.

[58] A. T. Özdemir, "An analysis on sensor locations of the human body for wearable fall detection devices: Principles and practice," *Sensors*, vol. 16, no. 8, p. 1161, 2016.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: arXiv:1412.6980.

[60] H. Wei, P. Chopada, and N. Kehtarnavaz, "C-MHAD: Continuous multimodal human action dataset of simultaneous video and inertial sensing," *Sensors*, vol. 20, no. 10, p. 2905, 2020.

Zhihang Zhong received the B.E. degree (Chu Kochen Hons.) from the School of Mechanical Engineering, Zhejiang University, Hangzhou, China in 2018, and the M.E. degree from the Graduate School of Engineering, The University of Tokyo, Tokyo, Japan, in 2020, where he is currently pursuing the Ph.D. degree with the Graduate School of Information Science and Technology.

His current research interests include computer vision, deep learning, and human–computer interaction.

Chingszu Lin received the B.E. degree in mechanical and electromechanical engineering from National Sun Yat-sen University, Kaohsiung, Taiwan, in 2012, and the M.E. and Ph.D. degrees from the Department of Precision Engineering, University of Tokyo, Tokyo, Japan in 2020.

His research interests include robot patients and humanoid robots.

Masako Kanai-Pak received the B.S.N. degree from Southern Oregon State College, Ashland, OR, USA, in 1985, the M.S.N. degree from the University of Hawaii at Manoa, Honolulu, HI, USA, in 1988, and the Ph.D. degree from the University of Arizona, Tucson, AZ, USA, in 2009.

She is a Professor with the College of Nursing, Kanto Gakuin University, Yokohama, Japan. She worked as a registered nurse in Japan. Her main research areas are leadership and management in nursing. She has been a Nursing Educator in Japan for over 30 years.

Prof. Kanai-Pak was a Board Member of the International Council of Nursing from 2009 to 2017.

Jukai Maeda received the Ph.D. degree in nursing from the Nagano College of Nursing, Komagane, Japan, in 2004.

Since 2009, he has been a Professor with Tokyo Ariake University of Medical and Health Sciences, Tokyo, Japan. He obtained an RN license with the University of Tokyo, Tokyo in 1989. After graduation, he worked with the management unit of Sony Corporation, Tokyo, for five years, following which he returned to the nursing world. He worked with the Nagano College of Nursing for 14 years as an Educator. One of his research interests is information processing in nursing.

Yasuko Kitajima received the bachelor's degree in economics from Senshu University, Tokyo, Japan, in 1999, the master's degree in economics from Saitama University, Saitama, Japan, in 2002, and the Doctoral degree in economics from the Graduate School of Economic Science, Saitama University in 2013.

She is also a qualified nurse and licensed emergency medical technician paramedic in Japan. She teaches Adult Nursing with Tokyo Ariake University of Medical and Health Sciences, Tokyo.

Mitsuhiro Nakamura received the B.S.N. and M.S.N. degrees from the Nagano College of Nursing, Komagane, Japan, in 2000 and 2006, respectively.

He is currently a Faculty Member of Tokyo Ariake University of Medical and Health Sciences, Tokyo, Japan. His research interests include nursing ethics and fundamental nursing.

Noriaki Kuwahara received the Dr.Eng. degree from the University of Tokyo, Tokyo, Japan, in 1996.

He joined Sumitomo Electric Industry, Ltd., Osaka, Japan, in 1987; ATR Communication System Labs, Kyoto, Japan, in 1993; and Kyoto Institute of Technology (KIT), Kyoto, Japan, in 2007. Since 2016, he has been a Professor of KIT.

Prof. Kuwahara is a member of the Society of Serviceology, Human Interface Society of Japan, the Institute of Image Information and Television Engineers, Japan Ergonomics Society, and the Textile Machinery Society Japan.

Taiki Ogata received the B.E., M.E., and Ph.D. degrees from the School of Engineering, University of Tokyo, Tokyo, Japan, in 2004, 2006, and 2009, respectively.

From 2009 to 2011, he was a Project Researcher with the Intelligent Modeling Laboratory, University of Tokyo. From 2011 to 2018, he was a Research Associate with Research into Artifacts, Center for Engineering, University of Tokyo. Since 2018, he has been a specially appointed Associate Professor with Tokyo Institute of Technology, Tokyo. His research interests include cognitive science, human–machine interaction, and human communication.

Dr. Ogata is a member of the Japan Society for Precision Engineering, the Society of Instrument and Control Engineers, and Society for Serviceology.

Jun Ota (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from the Faculty of Engineering, University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1994, respectively.

He is a Professor with Research into Artifacts, Center for Engineering, School of Engineering, University of Tokyo. From 1989 to 1991, he worked with Nippon Steel Corporation, Tokyo. In 1991, he was a Research Associate with the University of Tokyo. He became a Lecturer and an Associate Professor in 1994 and 1996, respectively. In April 2009, he became a Professor with the Graduate School of Engineering, University of Tokyo. In June 2009, he became a Professor with Research into Artifacts, Center for Engineering (RACE), University of Tokyo. Since 2019, he has been a Professor with RACE, School of Engineering, University of Tokyo. Since 2015, he has been a Guest Professor with the South China University of Technology, Guangzhou, China. From 1996 to 1997, he was a Visiting Scholar with Stanford University, Stanford, CA, USA. His research interests include multiagent robotic systems, design support for large-scale production/material handling systems, the science of hyperadaptability, and human behavior analysis/support.

Prof. Ota received a fellowship from the Robotics Society of Japan in 2016.