Deep Learning for Recognizing Human Activities using Motions of Skeletal Joints

Cho Nilar Phyo, *Student Member*, *IEEE*, Thi Thi Zin, *Member*, *IEEE* and Pyke Tin

***Abstract*— With advances in consumer electronics, demands have increased for greater granularity in differentiating and analyzing human daily activities. Moreover, the potential of machine learning, and especially deep learning, has become apparent as research proceeds in applications such as monitoring the elderly, and surveillance for detection of suspicious people and objects left in public places. Although some techniques have been developed for Human Action Recognition (HAR) using wearable sensors, these devices can place unnecessary mental and physical discomfort on people, especially children and the elderly. Therefore, research has focused on image-based HAR, placing it on the front line of developments in consumer electronics. This paper proposes an intelligent human action recognition system which can automatically recognize the human daily activities from depth sensors using human skeleton information, combining the techniques of image processing and deep learning. Moreover, due to low computational cost and high accuracy outcomes, an approach using skeleton information has proven very promising, and can be utilized without any restrictions on environments or domain structures. Therefore, this paper discusses the development of an effective skeleton information based HAR which can be used as an embedded system. The experiments are performed using two famous public datasets of human daily activities. According to the experimental results, the proposed system outperforms other state-of-the-art methods on both datasets.**

***Index Terms*— human action recognition, consumer electronics perspective, skeletal joints, formation of relative joint image, deep learning**

I. Introduction

HUMAN Action Recognition (HAR) from a set of video sequences is a challenging problem in computer vision technology, and is fundamental to a variety of applications in many different research areas, such as academia, security, industry, and consumer electronics. Among these are HAR systems used in video surveillance, consumer behavior analysis, and smart, in-home health-care monitoring systems for the elderly. HAR is an important research because a lot of potential accidents can be avoided by recognizing and predicting the activities of human being. On the other hand, recent advances in imaging technologies for consumer electronics, such as consumer depth sensors, have drawn a tremendous amount of attention from researchers in developing a variety of applications for the recognition and identification of human actions. However, modeling HAR system to produce accurate and efficient outcomes remains a challenge because of variations in scale, deformation and appearances. This paper proposes a new approach for establishing HAR in the consumer-electronics world by utilizing Color Skeleton Motion History Image (Color Skl-MHI) and Relative Joint Image (RJI) to monitor elderly people living alone.

As a consideration for developing applications in consumer electronics, the world's population is aging. In the future, the elderly are expected to comprise a greater proportion of the total population. In addition, most of the elderly prefer to live independently in their homes in a familiar environment. This presents a challenge to caregivers in trying to provide quality services. At the same time, affordable services are hard to find for many of the elderly. Therefore, automatic systems for monitoring the daily activities of the elderly have become an important subject of research in consumer electronics, specifically the automatic recognition of human actions and interactions with consumer electronics such as smartphones, as well as home appliances, medicines and many other objects. This paper deals with this subject by focusing on a hybrid technology of image processing and deep learning to establish a human action recognition system in consumer electronics. This paper discusses improvements in three areas.

- In methodology, this paper proposes using a depth sensor for recognizing human daily activities by tracking the motions of skeletal joints. These techniques can be applied to the analysis of various types of human action and interaction in various environments both day and night.
- In application, this proposed system is very useful for monitoring elderly people living alone by recognizing and analyzing the normal and abnormal daily activities of elderly.
- In implementation, this proposed system can be developed as an embedded system for real-time processing.

This paper is organized as follows: Part II presents other work related to the theme of this paper; Part III provides an overview and technical explanations of the proposed human action recognition; Part IV shows some experimental results from two public datasets; Part V provides analysis and discussion. Finally, Part VI provides a conclusion.

This work was supported in part by the Telecommunication Advanced Foundation.

C. N. Phyo is with Interdisciplinary Graduate School of Agriculture and Engineering, University of Miyazaki, Miyazaki, Japan (e-mail: nc16004@student.miyazaki-u.ac.jp).

T. T. Zin is with Graduate School of Engineering, University of Miyazaki, Miyazaki, Japan (e-mail: thithi@cc.miyazaki-u.ac.jp).

P. Tin is with International Relation Center, University of Miyazaki, Miyazaki, Japan (email: pyketin11@gmail.com).

II. Some Related Works

The research area of Human Action Recognition (HAR) is related to many different research fields where successful applications have been developed using HAR as a base in consumer electronics. In daily life, human beings act or interact with devices such as smartphones, other consumer electronics, home appliances, and many other objects. HAR research has been conducted in the framework of consumer electronics, health-care monitoring systems, video surveillance and deep learning technology. The following describes some of the many excellent studies.

A. HAR and Consumer Electronics

Human daily activities very much involve acting and interacting with consumer electronics devices such as smartphones, home appliances, and many other objects. Therefore, research involving such activities plays a key role in developing HAR for consumer electronics technologies [1]. Human beings are connected with millions of consumer electronics devices in the Internet of Things. On the other hand, side effects might exist. For example, many people use computers in home and office, resulting in increasing lengths of time sitting in one place, often with poor posture. This can lead to repetitive stress injuries. Thus, it is useful to monitor daily routines, determine unhealthy behavior, and take appropriate actions. Using self-collected datasets, such problems have been modeled, including illustrations, and both hardware and software technologies have been developed for recognizing human activities [2].

Similarly, a system using a consumer video sensor and a Hidden Markov Model has been developed for recognizing six abnormal activities of the elderly, such as falling forward, falling backward, fainting, vomiting, and experiencing headaches or chest pain [3]. Alternatively, a consumer depth camera was utilized to capture depth images of six typical human actions such as walking, running, boxing, clapping, sitting down, and standing-up. Then, a method of recognizing those activities was developed using the R-transform method for feature extraction, Principle Component Analysis (PCA) for reducing the dimensions of feature vectors, Linear Discriminant Analysis (LDA) for extracting more prominent features, and the Hidden Markov Model for classification [4].

B. HAR and Health Care Monitoring

HAR is an important part of establishing an intelligent health-care monitoring system, especially for allowing the elderly to live independently with a high quality of life. It is also important to establish the regularity and normal timing of activities such as taking medicines and meals to help develop guidelines and regulations. Among the many studies, some have used a multilayered Markov Chain model [5] and a Markov based approach for recognizing daily living activities [6] utilizing image technology to analyze and recognize not only simple actions such as walking, standing, and sitting, but also compound activities such as doing housework, communicating, and taking medicine.

Apart from image technologies, much research has focused on human action recognition and monitoring using wearable consumer devices for measuring physiological electrocardiogram (ECG) signals and using Global Positioning System (GPS) for finding an elderly person who has fallen in the outdoors environment [7]. Another fall-detection system has been developed by applying sensors within the home network [8]. As another aspect of research, health-management systems have been developed for the elderly using the Radio Frequency Identification (RFID), short-range wireless transmission technology, and the internet for blood pressure management [9].

C. HAR and Video Surveillance

Video surveillance is another area of application-oriented HAR research. For example, video surveillance has been used in public places such as shopping centers, sports centers and transportation stations to detect suspicious people and suspicious objects. An intelligent video surveillance system has been developed that can automatically detect loitering people using the two-dimensional Random Walk Model [10]. In consumer electronics research, as part of studying human interaction with objects, abandoned object detection was investigated by developing a stationary object analyzer [11].

D. HAR and Deep Learning

HAR and machine learning in general, as well as deep learning in particular, can be considered complementary areas of research. As a demonstration of this statement, a two-step labeling scheme that utilizes deep learning technology over spatial-temporal features in a grayscale image was developed to recognize human actions in research by Baccouche, M., et al. [12]. Alternatively, different approaches for handling HAR have been attempted using other types of deep learning, such as 3D-based Deep Convolutional Neural Networks ($3D^2CNN$) and Support Vector Machine (SVM). In one approach, $3D^2CNN$ was directly applied on raw depth video sequences for extracting spatial-temporal features. In another, SVM was utilized over joint vector features, which were based on simple position and angle information between human skeletal joints [13]. In this work, the results of $3D^2CNN$ and SVM were fused to produce the final output results for recognizing action. In other work on HAR with deep learning, the 3D CNN model has been used by adding a hardwired layer before the first convolution layer, thus generating five channels of information which were used in the final layer for recognizing actions [14]. It is also worthwhile to mention other, more relevant work in which the concept of view-invariant HAR was illustrated by applying enhanced skeleton visualization and experimentation. This work used Northwestern-UCLA, UWA3DII, NTU RGB+D and MSRC-12 datasets [15]-[16]. Last but not least in related work, some research has been attempted to overcome limitations of initial camera angles and positions by developing an algorithm for calibrating skeleton coordinates [17].

This has been an overview of related HAR research and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCE.2019.2908986, IEEE Transactions on Consumer Electronics

application in the consumer electronics world. Actually, the literature provides numerous examples of such research, with a variety of potential applications. However, none of this work has developed technology with a satisfactory level of performance. Therefore, the following new approach to HAR has been developed.

III. Proposed Human Action Recognition

The proposed HAR architecture is described in Fig. 1. The system contains three processing components: 1) input data acquisition, 2) the creation of sub-images and feature extraction, and 3) output fusion and action recognition. In input data acquisition, the input device captures coordinate information for the joints of the human body. In the component involving images and features, (i) a Color Skeleton Motion History Image (Color Skl-MHI) is created for extracting motion history features, and (ii) a Relative Joint Image (RJI) is created for obtaining the relative distance features between the referenced joints and the other joints. In the component for output fusion and human action recognition, deep learning is performed, as well as classification into one of the predefined actions such as standing, sitting, or bending. Fig. 2 provides an illustrative application of the HAR system using some consumer electronics products for monitoring the elderly living alone. In the following, the functions of each component are presented.

A. Function of Input Data Acquisition Component

Input data acquisition is done using a depth sensor that can generate 5 kinds of output streams: RGB, depth, infrared, audio, and joints tracking data. These data are very useful for recognizing human actions and interactions. However, only joints tracking data are used for the implementation of the proposed HAR system. The reason is that in the skeleton tracking data, the depth sensor can track the motion of the human body and generate joint coordinates which include 3D coordinate of $J_i(t) = [x_i(t), y_i(t), z_i(t)]$ at a frame in time t. This joint information represents the structure of human body quite well.

B. Creation of Sub-images and Feature Extraction

1) Creating Color Skeleton Motion History Image

In this component, Color Skl-MHI and RJI are established using joints information that is obtained using the depth sensor. Creating a Skeleton Motion History Image (Skl-MHI), firstly involves creating a binary skeleton image using the corresponding joints. Next, a Binary Skl-MHI is created by combining the continuous skeleton images within the predefined temporal dimension [18]. However, the Binary Skl-MHI alone cannot differentiate actions that have a similar motion pattern, such as sitting down from a standing position, or standing up from a sitting position.

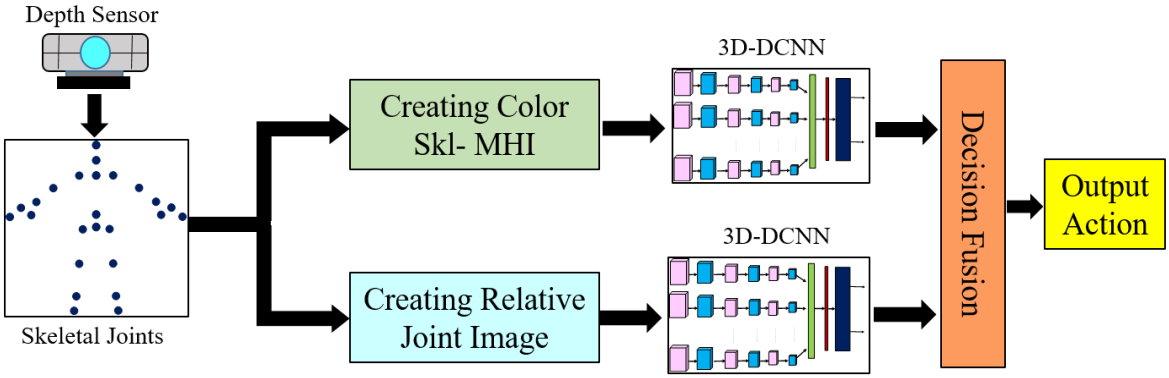

Fig. 1. Architecture of the Proposed Human Action Recognition System

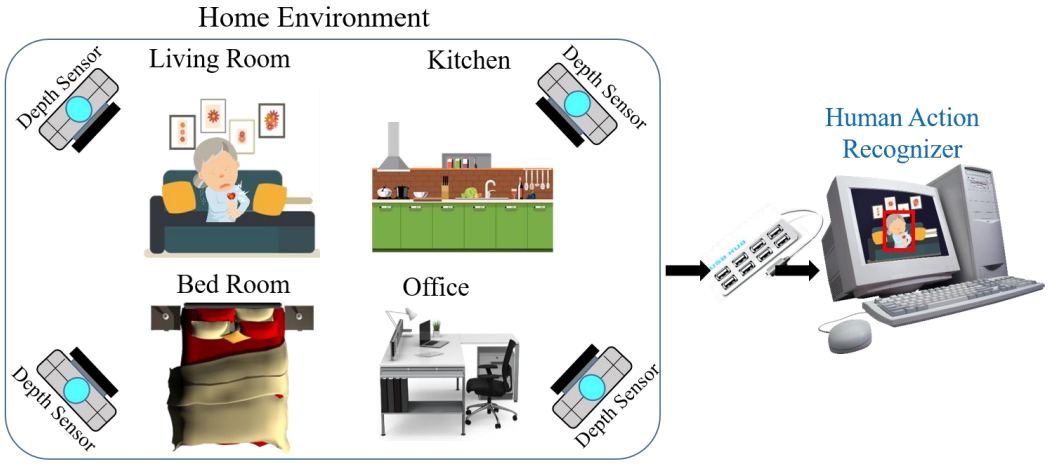

Fig. 2. Consumer HAR for Monitoring Eldery People Living Alone

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCE.2019.2908986, IEEE Transactions on Consumer Electronics

Therefore, the Binary Skl-MHI is added with the color value according to the sequential time interval of each action as shown in Fig. 3. Consequently, the time sequence of motion history images derived from skeleton data becomes more obvious. The outputs of this process are shown in Fig. 4. In this figure, we can see that applying a color value for each time sequence representation, enables differentiating the actions of very similar motion patterns.

2) *Creating Relative Joint Image*

The use of the relative positions of joints is an intuitive method of representing human motion. For example, in detecting the action of waving the hands, hand positions are located above shoulder joints and move in left and right directions. The RJI is the transformation of the skeleton sequence of an action into a new view-invariant representation of a human motion pattern image. For creating RJI, firstly we calculate the relative distance between referenced joints and the other joints. We use as referenced joints the four joints of left shoulder (J5), the right shoulder (J9), the left hip (J13) and the right hip (J17), because they are the most stable joints in most actions.

Then, the relative distance features are extracted by subtracting (x, y, z) coordinates between four referenced joints and the other joints. By contacting the relative distance of the joints in the frames within the predefined temporal dimension, four RJI, each containing the 3D array with the size of $t \times (m\text{-}1) \times 3$ are generated as shown in Fig. 5, where m is the number of skeletal joints in each frame, t is the predefined temporal dimension that includes enough frames for representing the action, and 3 is for (x, y, z) coordinates.

C. *Outputs Fusion and Human Action Recognition*

In this component, three dimensional deep convolutional neural networks (3D-DCNN) is applied to Color Skl-MHI and RJI for training and recognizing the human actions. The final output action is generated by fusing the results of those two networks. The deep convolutional neural network (DCNN) is a multi-layer convolutional neural network which consists of an input layer, one or more hidden layers, a fully-connected layer, and an output layer. The hidden layer of the DCNN contains two operations, namely convolution and pooling, and one main function called the activation function.

In the output layer, the soft-max function is applied for transforming the output of DCNN into the corresponding probability value. The weight values for each convolution kernel is randomly initialized using some distribution concepts such as Uniform distribution and Gaussian distribution. The network is trained forward (feeding inputs data along the network) and backward (updating the weights of all layers) multiple times, until it meets the stopping criteria (minimum loss or maximum iteration). The weight values are updated depending on the loss value which is calculated using the stochastic gradient descent algorithm.

1) *Convolution*

Convolution is the mixing of information that is obtained from the raw input data and the extracted representative data that can well represent the characteristics of the input training data. The output of the convolution is called a feature map, and it can differ, depending on the kernel filter that is used for the convolution operation. When convolution is applied over the image data, a 2D kernel is utilized on each color channel of the image. Fig. 6 (a) shows the example of applying a convolution operation on the image using the edge detection kernel.

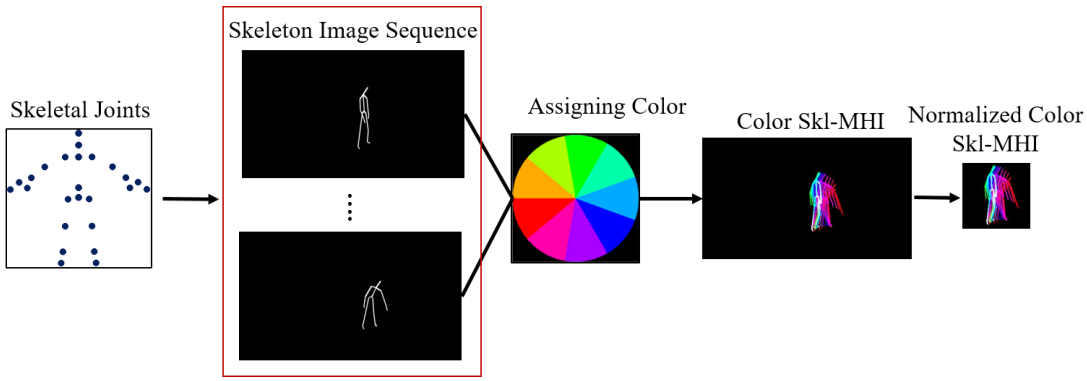

Fig. 3. Process of Creating Color Skeleton Motion History Image (Skl-MHI)

Fig. 4. Binary Skl-MHI and Color Skl-MHI for Actions with Similar Motion Patterns: (a) Sitting down, (b) Standing up

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCE.2019.2908986, IEEE Transactions on Consumer Electronics

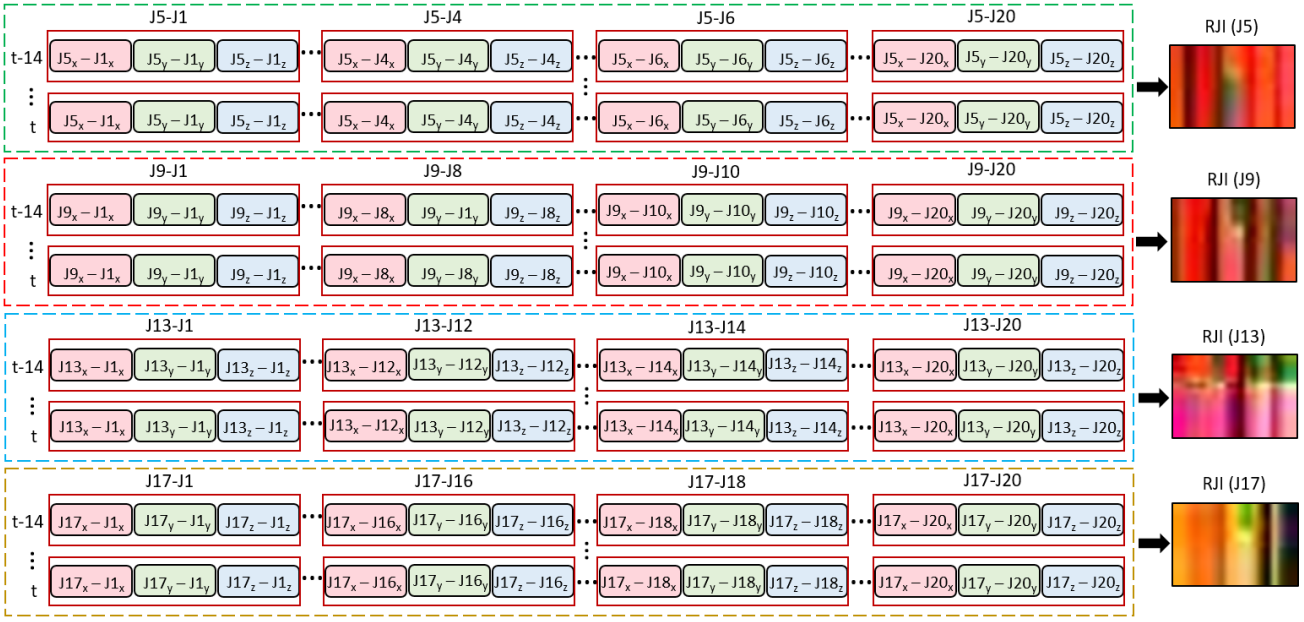

Fig. 5. Creating Images of Relative Positions for Four Referenced Joints (J5, J9, J13 and J17)

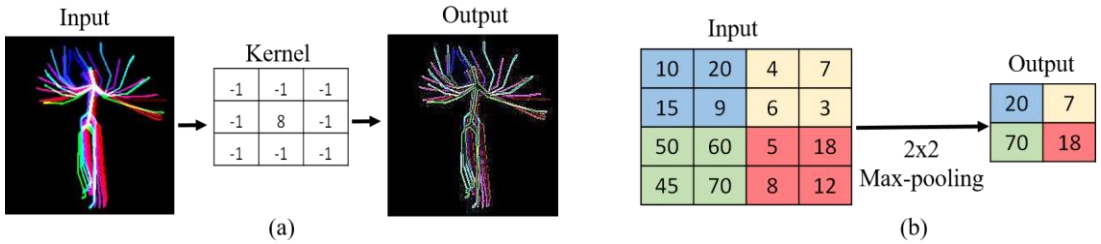

Fig. 6. Example of Hidden Layer Operations of DCNN (a) Convolution Using Edge Detection Kernel, (b) Illustration of Max-pooling Operation

2) Pooling

The main purpose of the pooling operation is to reduce the spatial dimension of the feature map in each hidden layer. There are several types of pooling, depending on the mathematical operation that is used for performing the pooling process, such as maximum pooling (MAX), and average pooling (AVE). The pooling operation is also known as sub-sampling, which is a simple operation of taking the maximum or average value within the predefined kernel's width and height. Fig. 6 (b) shows the example of applying a maximum pooling operation over a 2D matrix.

3) Activation Function

The activation function is an important function in DCNN, because it can decide whether the neuron should be activated or not, depending on whether the received input is relevant. The output of the activation function becomes input for the next layer, and the data for the multiplication of the neuron's input with the corresponding kernel weight and addition with bias value is used as input for the activation function which can be expressed as (1).

$$Y = f\left(\sum(\text{kernel\_weight} * \text{input}) + \text{bias}\right) \quad (1)$$

In this proposed system, we use the activation function of the Rectified Linear Unit (ReLU), which produces 0 when $x < 0$, and is linear with slope 1 when $x > 0$, as described in (2).

$$f(x) = \max(0, x) \quad (2)$$

where x is the input of the activation function.

4) Soft-Max Function

The soft-max function is applied after the output layer of 3D-DCNN in order to obtain the probability of the possible action using (3).

$$\sigma(z)_j = e^{z_j} / \sum_{k=1}^{N} e^{-z_k} \quad (3)$$

where j and z denotes each action and its network output and N represents the total number of actions.

5) Updating Weights using Stochastic Gradient Descent Method

The step-by-step calculations for updating the weights of the kernel matrix of all layers using the stochastic gradient descent algorithm [19] is as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCE.2019.2908986, IEEE Transactions on Consumer Electronics

i. The error value err between desired output $y_{d,k}$ and network output y_k is calculated using (4).

$$err = y_{d,k} - y_k \tag{4}$$

ii. The error gradient for neurons in the output layer Δ_k is calculated using (5).

$$\Delta_k = y_k(1-y_k)err \tag{5}$$

iii. The weight and bias correction for output layer ΔW_k and ΔB_k are calculated using a predefined learning rate of α, as described in (6) and (7).

$$\Delta W_k = \alpha y_{hk} \Delta_k \tag{6}$$

$$\Delta B_k = \alpha(-1)\Delta_k \tag{7}$$

where y_{hk} represents the hidden layer output.

iv. The error gradient for neurons in the hidden layer Δ_{hk} is calculated using (8).

$$\Delta_{hk} = y_{hk}(1-y_{hk})\Delta_k \Delta W_k \tag{8}$$

v. The weight and bias correction values for the hidden layers ΔW_{hk} and ΔB_{hk} are calculated using predefined learning rate α, as described in (9) and (10).

$$\Delta W_{hk} = \alpha y_{hk} \Delta_{hk} \tag{9}$$

$$\Delta B_{hk} = \alpha(-1)\Delta_{hk} \tag{10}$$

vi. The weights and bias of all layers $W_{i,new}$ and $B_{i,new}$ except the input layer are updated using the corresponding weight and bias correction values as described in (11) and (12).

$$W_{i,new} = W_{i,old} + \Delta W_i \tag{11}$$

$$B_{i,new} = B_{i,old} + \Delta B_i \tag{12}$$

vii. Step (i) through (vi) are repeated until stopping criteria are met.

6) *Architecture of 3D-DCNN*

The 3D-DCNN architecture for training the Color Skl-MHI and RJI data for human activity is shown in Fig. 7. In this architecture, normalized Color Skl-MHI of size 62×62 and RJI of size 15×19 are used as input data and each hidden layer is composed of the operations of convolution (Conv) and pooling (Pool), dropout (Drop), and neuron activation (ReLU). This architecture includes three hidden layers (i =1, 2, 3). The predefined kernel size and initialization type of each Conv and Pool layers are described in Table Ⅰ. The dropout ratios for the three hidden layers are 0.1%, 0.2%, and 0.3%, respectively.

For the weight initialization for the Conv layer, the method developed by Microsoft Research Asia (MSRA) is used. This is well-suited for ReLU. For the fully connected layer (FC), a weight vector of size 1,000 is used, and a soft-max function is applied in the output layer to obtain the probability for each possible action. All weights are trained using the back propagation algorithm with stochastic gradient descent method.

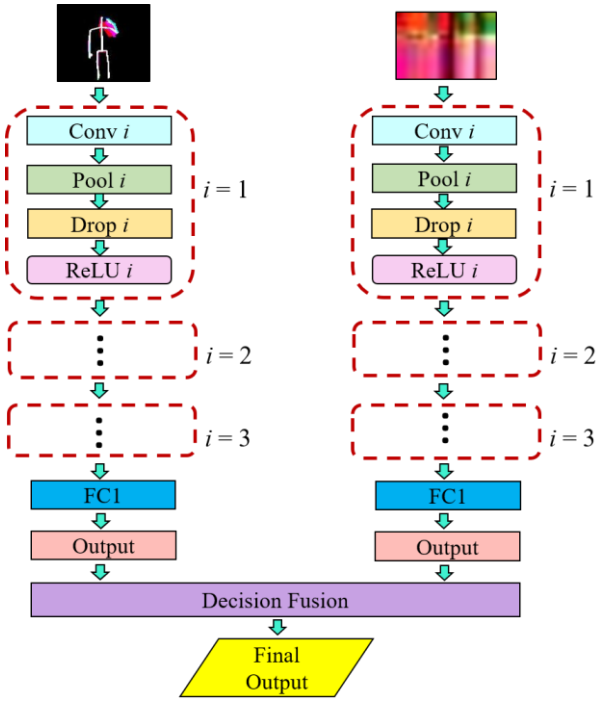

Fig. 7. Architecture of Proposed 3D-DCNN

TABLE I
PARAMETERS OF CONVOLUTION AND POOLING LAYERS

Layers	Filter Size	Initialization Type	Feature Maps
Conv 1	7×7	MSRA	3
Pool 1	2×2	MAX	3
Conv 2	5×5	MSRA	10
Pool 2	2×2	AVE	10
Conv 3	3×3	MSRA	15
Pool 3	2×2	AVE	15

IV. EXPERIMENTAL RESULTS

We performed experiments on the proposed HAR under various conditions using two different datasets: the UTKinect Action-3D dataset, and the CAD-60 state-of-the-art public human-activity 3D dataset. These datasets included daily activities such as drinking water, answering the phone, and cooking. In the UTKinect Action-3D dataset, we viewed the same actions using several camera view-points and time durations. In this dataset, we also viewed the same actions performed by different people, each performing the action in a different way. This is called high intra-class variation. In the CAD-60 dataset, we performed these experiments in several different environments, such as kitchens, offices and bath rooms. This dataset includes a high degree of similarity between different actions, which is particularly challenging for recognition technology. The experimental results from these two datasets show the effectiveness of proposed HAR. The detail parameters and experimental data can be downloaded at *http://github.com/CNPhyo/ColorSklMHI-and-RJI-based-HAR.*

A. *UTKinect Action-3D Dataset*

The videos in this dataset are captured using a single stationary-depth camera, and consist of 10 actions performed by 10 different subjects. Each subject performs all actions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCE.2019.2908986, IEEE Transactions on Consumer Electronics

twice. The dataset provides the 3D locations of 20 joints in the 199 action sequences [17]. This dataset includes variations in viewpoint and high intra-class variations. For performing the experiments, we create the Color Skl-MHI and RJI data by connecting 20 skeletal joints and combining them within every 15 frames. For the performance evaluation, a cross-validation method is used involving the omission of one subject. We train 3D-DCNN models by alternatively using the actions of 9 subjects as training data and omitting 1 subject's data for testing. In the case of training 3D-DCNN using RJI data, it needs to train 3D-DCNN models for 4 RJIs: RJI (J5), RJI (J9), RJI (J13), RJI (J17). Therefore, the total trained models using RJI data is 40 (4 × 10 subjects) and Color Skl-MHI based model is 10 (1 × 10 subjects) for all experiments. Firstly, 3D-DCNN is trained using Color Skl-MHI. The sample Color Skl-MHI of 10 daily activities of UTKinect Action-3D dataset are shown in Fig. 8. In the evaluation, the Color Skl-MHI based method achieves an overall accuracy of 94%. Then, 3D-DCNN is also trained for RJI. The sample RJI for four referenced joints (J5, J9, J13, J17) involved in the actions of picking up objects and waving hands are shown in Fig. 9 and Fig. 10. With the same evaluation method, RJI based 3D-DCNN achieves an overall accuracy of 95%. After training the 3D-DCNN for both Color Skl-MHI and RJI, the decision of those two 3D-DCNN are fused by averaging the probability of the corresponding actions and taking the action which has maximum probability. The detailed confusion matrix for the decision fusion approach is described in Fig. 11. In Table II, we can see that proposed system achieved an overall accuracy of 97%, which is higher than other state-of-the-art methods.

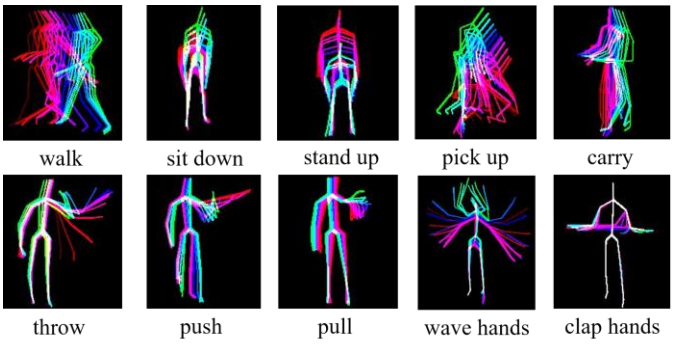

Fig. 8. Color Skl-MHI for 10 Daily Activities from UTKinect Action-3D

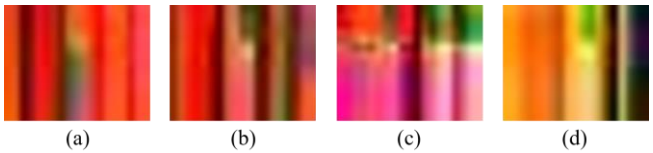

Fig. 9. Images of Relative Joint Positions for Picking up Objects, (a) RJI (J5), (b) RJI (J9), (c) RJI (J13), (d) RJI (J17)

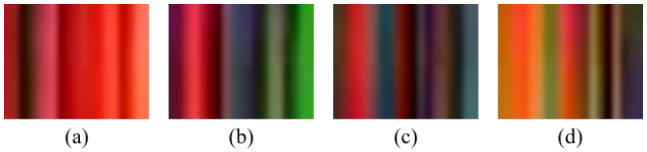

Fig. 10. Images of Relative Joint Positions for Waving Hands, (a) RJI (J5), (b) RJI (J9), (c) RJI (J13), (d) RJI (J17)

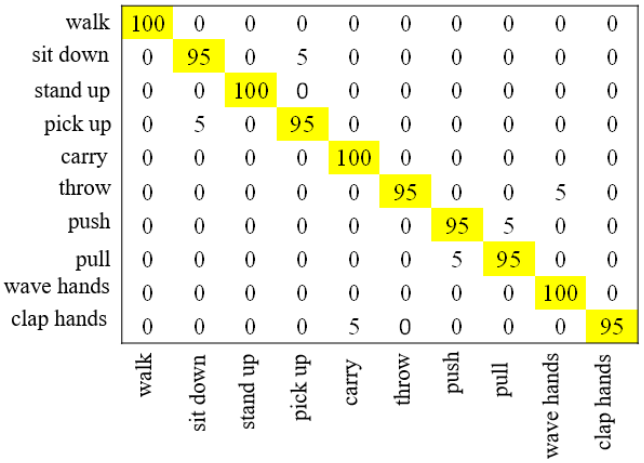

TABLE II
RESULTS ON THE UTKINECT ACTION-3D DATASET

Method	Accuracy (%)
L. Mengyuan *et al.* (2016) [15]	95.50
L. Zhi *et al.* (2014) [16]	95.00
X. Lu *et al.* (2012) [17]	90.92
M. Devanne *et al.* (2013) [21]	91.50
A. Chrungoo *et al.* (2014) [22]	91.96
Color Skl-MHI 3D-DCNN	94.00
RJI 3D-DCNN	95.00
Proposed System	97.00

B. CAD-60 Daily Activity Dataset

The CAD-60 daily activity dataset contains 12 high-level daily activities and 1 still activity performed by 4 people in 5 different environments, such as office, kitchen, bedroom, bathroom, and living room [23]. This dataset is challenging because of the high similarity between actions. In the creation of Color Skl-MHI and RJI data, we use the skeletal joints within every 15 frames. For improving the performance of recognition for a left-handed person, Color Skl-MHI and RJI data are mirrored in order to make her activities similar to the other three right-handed people. The sample Color Skl-MHI for 10 daily activities in the CAD-60 dataset are shown in Fig. 12. For the performance evaluation, we use the same method that was used on the UTKinect Action-3D dataset. We train 3D-DCNN models by alternatively using 3 people' data for training and 1 omitted person data for testing. The total trained models using Color Skl-MHI is 4 (1 × 4 people) and RJI is 16 (4 × 4 people) for all experiments. On CAD-60 daily activity dataset, Color Skl-MHI and RJI based 3D-DCNN achieves an overall accuracy of 71.15% and 86.5%, respectively, and the decision fusion approach achieved 92.31%.

In Fig. 13, we can see that the proposed system can recognize most of the actions with 100% accuracy. However, a few actions are misrecognized because of the proximity of hand locations. The environment was taken into consideration to overcome such problems. After applying these environmental considerations, the overall accuracy increased to 96.15%, as shown in Fig. 14. The failures to differentiate the actions of brushing teeth and talking on the phone, as well as brushing teeth and drinking water, were eliminated because these actions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCE.2019.2908986, IEEE Transactions on Consumer Electronics

are performed in different environments. Table Ⅲ compares the accuracy, precision and recall of other state-of-the-art methods. As we can see, the proposed system provided superior performance. The results of the other methods are taken from the CAD-60 daily activity dataset website. This website includes reports on performance of the other methods using this dataset [23]. Because the other authors reported the performance of their algorithm in terms of either precision and recall or accuracy, Table Ⅲ has some blank cells.

V. Discussion

The proposed system has low computational cost because of its use of Color Skl-MHI and RJI. The results are very promising for real-time applications. The processing time of the proposed system is evaluated based on the feature extraction time of Color Skl-MHI and RJI, as well as the classification time using 15 fps video data. For comparing time complexity, the proposed system is tested in the same hardware environment as described in [17]. Then, our proposed system takes 0.0636 s for feature extraction, and has an average testing time of 0.0081 s for 1 input sequence from the UTKinect Action-3D dataset while the system in [17] has an average testing time of 0.0125 s for a sequence from the same dataset. Moreover, the average time consumed for feature extraction using our proposed system is 0.1179 s and the method in [32] consumes 0.18 s for 1 sequence from UTKinect Action-3D dataset when executing on a machine with same specification that used in [32].

In addition, because of the use of Color Skl-MHI, actions from the UTKinect Action-3D dataset with a similar motion pattern, such as throwing and pushing, were correctly differentiated as shown in Fig. 15 (a) and Fig. 15 (b). From the CAD-60 daily activity dataset, the proposed system misrecognized actions such as drinking water and talking on the phone. As shown in Fig. 16 (a) and Fig. 16 (b), this happened because those actions contain very few motions and the hands are very close. This points out the fact that we have to consider history of previous states for correctly classifying those actions.

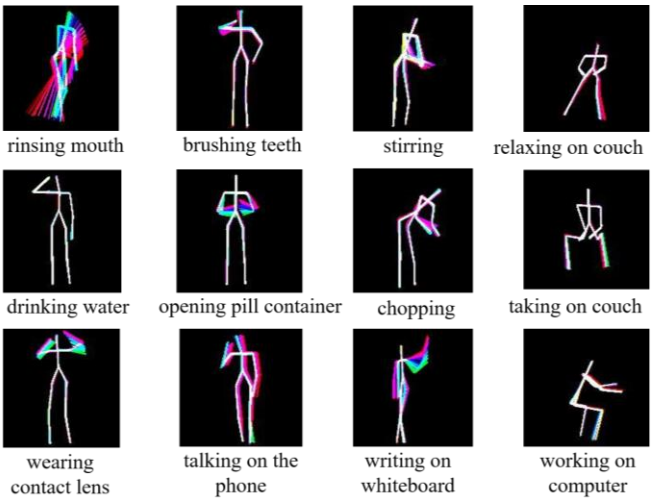

Fig. 12. Color Skl-MHI for 12 Daily Activities from the CAD-60 Dataset

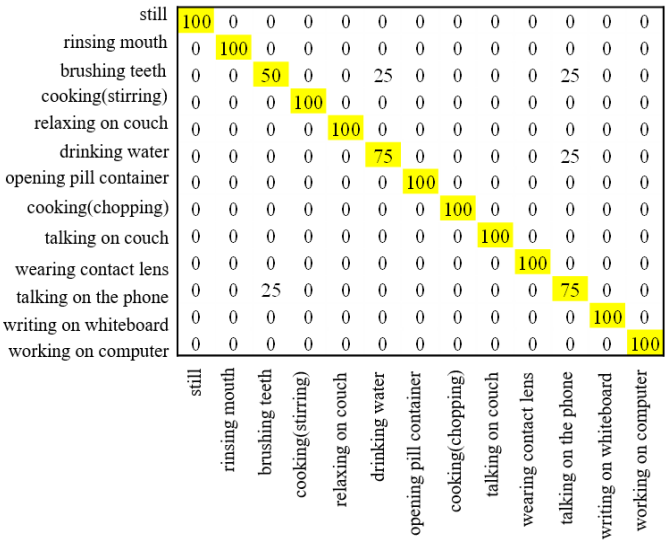

Fig. 13. Confusion Matrix on the CAD-60 Dataset

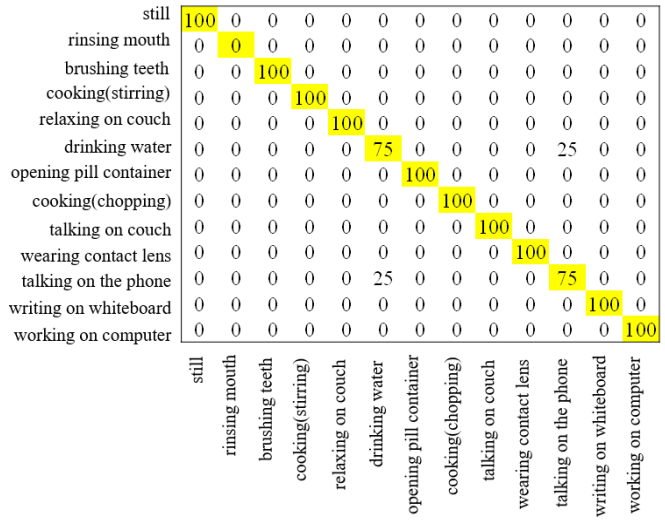

Fig. 14. Confusion Matrix on the CAD-60 Dataset with Environment

TABLE III
Results on the CAD-60 Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)
N. Bingbing *et al.* (2012) [24]	65.32	-	-
R. Gupta *et al.* (2013) [25]	-	78.10	75.40
J. Wang *et al.* (2013) [26]	74.70	-	-
Y. Zhu *et al.* (2014) [27]	-	93.20	84.60
D. R. Faria *et al.* (2014) [28]	-	91.10	91.90
J. Shan *et al.* (2014) [29]	-	93.80	94.50
S. Gaglio *et al.* (2014) [30]	-	77.30	76.70
G. I. Parisi *et al.* (2015) [31]	-	91.90	90.20
E. Cippitelli *et al.* (2016) [32]	-	93.90	93.50
Color Skl-MHI 3D-DCNN	71.15	70.02	69.39
RJI 3D-DCNN	86.50	82.78	82.08
Proposed System	96.15	90.39	88.46

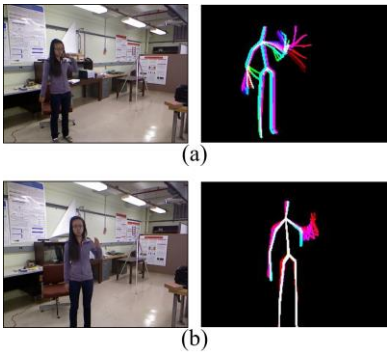

Fig. 15. Some Sample Actions and Color Skl-MHI of the UTKinect Action-3D Dataset (a) Throwing, (b) Pushing

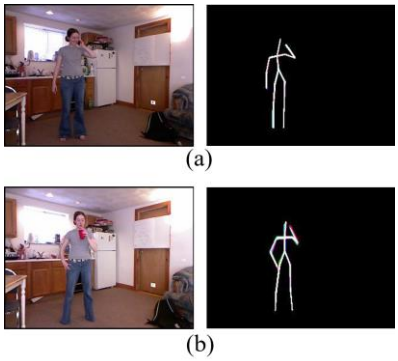

Fig. 16. Some Sample Actions and Color Skl-MHI from the CAD-60 Dataset (a) Talking on the Phone, (b) Drinking Water

VI. Conclusion

In this paper, we propose an intelligent human action recognition system to develop as a consumer electronics product (with low computational cost and high accuracy outcomes) for automatically monitoring and recognizing the daily activities of elderly people living alone. Moreover, this system can be utilized without any restrictions on environmental conditions or domain structures, and is also very promising for real-time applications because of the fast processing time. The problems of view-variation (single camera) and intra-class variation have been solved in this system. The experimental results show that the proposed system is outperforming other state-of-the-art methods both on UTKinect Action-3D and CAD-60 daily activity datasets. In the future, this system will be improved by taking into consideration the affordance hypothesis in the recognition of actions. More experiments will be performed on the datasets of complex actions which are related to health-problems, such as headaches and vomiting.

References

[1] P. A. Sandy, "Automatic mapping and modeling of human networks", *Physica A: Statistical Mechanics and its Applications*, vol. 378, no. 1, pp. 59-67, May 2007, 10.1016/j.physa.2006.11.046.

[2] N. Yugo *et al.*, "SenStick: comprehensive sensing platform with an ultra-tiny all-in-one sensor board for IoT research", *Journal of Sensors*, vol. 2017, p. 1-16, Oct. 2017, 10.1155/2017/6308302.

[3] Z. A. Khan and W. Shon, "Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care", *IEEE Trans. Consumer Electronics*, vol. 57, no. 4, pp. 1843-1850, Nov. 2011, 10.1109/TCE.2011.6131162.

[4] A. Jalal *et al.*, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home", *IEEE Trans. Consumer Electronics*, vol. 58, no. 3, pp. 863-871, Sept. 2012, 10.1109/TCE.2012.6311329.

[5] T. T. Zin, P. Tin and H. Hama, "Visual monitoring system for elderly people daily living activity analysis", *in Proc. of the Int. MultiConf. of Engineers and Computer Scientists 2017*, Hong Kong, 15-17 Mar. 2017, pp. 140-142.

[6] L. Zaineb *et al.*, "A Markovian-based approach for daily living activities recognition", *in Proc. of the 5th Int. Conf. on Sensor Networks*, Rome, Italy, 17-19 Feb. 2016, pp. 214-219.

[7] L. H. Wang *et al.*, "An outdoor intelligent healthcare monitoring device for the elderly", *IEEE Trans. Consumer Electronics*, vol. 62, no. 2, pp. 128-135, Jul. 2016, 10.1109/TCE.2016.7514671.

[8] J. Wang *et al.*, "An enhanced fall detection system for elderly person monitoring using consumer home networks", *IEEE Trans. Consumer Electronics*, vol. 60, no. 1, pp. 23-29, Apr. 2014, 10.1109/TCE.2014.6780921.

[9] C. H. Hung *et al.*, "Design of blood pressure measurement with a health management system for the aged", *IEEE Trans. Consumer Electronics*, vol. 58, no. 2, pp. 619-625, Jul. 2012, 10.1109/TCE.2012.6227468.

[10] T. T. Zin, P. Tin, T. Toriu and H. Hama, "A Markov random walk model for loitering people detection", *in Proc. of the 6th Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, Darmstadt, Germany, 15-17 Oct. 2010, pp. 680-683.

[11] T. T. Zin, P. Tin, H. Hama and T. Toriu, "Unattended object intelligent analyzer for consumer video surveillance", *IEEE Trans. Consumer Electronics*, vol. 57, no. 2, pp. 549-557, Jul. 2011, 10.1109/TCE.2011.5955191.

[12] B. Moez *et al.*, "Sequential deep learning for human action recognition", *in Proc. of the Int. Workshop on Human Behavior Understanding*, Amsterdam, Netherlands, 16 Nov. 2011, pp. 29-39.

[13] L. Zhi *et al.*, "3D-based deep convolutional neural network for action recognition with depth sequences", *Image and Vision Computing*, vol. 55, no. 2, pp. 93-100, Nov. 2016, 10.1016/j.imavis.2016.04.004.

[14] J. Shuiwang *et al.*, "3D convolutional neural networks for human action recognition", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, Jan. 2013, 10.1109/TPAMI.2012.59.

[15] L. Mengyuan *et al.*, "Enhanced skeleton visualization for view invariant human action recognition", *Pattern Recognition*, vol. 68, pp. 346-362, Aug. 2017, 10.1016/j.patcog.2017.02.030.

[16] L. Zhi *et al.*, "An effective view and time-invariant action recognition method based on depth videos", *in Proc. of the IEEE Int. Conf. on Visual Communications and Image Processing*, Singapore, 13-16 Dec 2015, pp. 1-4.

[17] X. Lu *et al.*, "View invariant human action recognition using histograms of 3d joints", *in Proc. of the Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, 16-21 Jun. 2012, pp. 20-27. [Online]. Available: http://cvrc.ece.utexas.edu/KinectDatasets/HOJ3D.html.

[18] W. Tao, Y. Qiao and B. Lee, "Kinect skeleton coordinate calibration for remote physical training", *in Proc. of the 6th Int. Conf. on Advances in Multimedia*, Nice, France, 23-27 Feb. 2014, pp. 66-71.

[19] C. N. Phyo, T. T. Zin and P. Tin, "Skeleton motion history based human action recognition using deep learning", *in Proc. of 2017 IEEE 6th Global Conf. on Consumer Electronics*, Nagoya, Japan, 24-27 Oct. 2017, pp. 784-785.

[20] Y. Lecun *et al.*, "Gradient-based learning applied to document recognition," *in Proc. of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, 10.1109/5.726791.

[21] M. Devanne *et al.*, "Space-time pose representation for 3d human action recognition", *in Proc. of the Int. Conf. on Image Analysis and Processing*, Naples, Italy, 9-13 Sept. 2013, pp. 456-464.

[22] A. Chrungoo *et al.*, "Activity recognition for natural human robot interaction", *in Proc. of the Int. Conf. on Social Robotics*, Bristol, UK, 27-29 Oct. 2014, pp. 84-94.

[23] J. Sung *et al.*, "Unstructured human activity detection from rgbd images", *in Proc. of the IEEE Int. Conf. on Robotics and Automation*, Saint Paul, MN, USA, 14-18 May 2012, pp. 842-849. [Online]. Available: http://pr.cs.cornell.edu/humanactivities/data.php.

[24] N. Bingbing, P. Moulin and S. Yan, "Order-preserving sparse coding for sequence classification", *in Proc. of the European Conf. on Computer Vision*, Florence, Italy, 7-13 Oct. 2012, pp. 173-187.

[25] R. Gupta, A. Y. S. Chia and D. Rajan, "Human activities recognition using depth images", *in Proc. of the 21st ACM Int. Conf. on Multimedia*, Barcelona, Spain, 21-25 Oct. 2013, pp. 283-292.

[26] J. Wang *et al.*, "Learning actionlet ensemble for 3D human action recognition", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914-927, May 2014, 10.1109/TPAMI.2013.198.
[27] Y. Zhu, W. Chen and G. Guo, "Evaluating spatiotemporal interest point features for depth-based action recognition", *Image and Vision Computing*, vol. 32, no. 8, pp.453-464, Aug. 2014, 10.1016/j.imavis.2014.04.005.
[28] D. R. Faria, C. Premebida and U. Nunes, "A probalistic approach for human everyday activities recognition using body motion from RGB-D images", *in Proc. of the 23rd IEEE Int. Symposium on Robot and Human Interactive Communication*, Heriot-Watt University, Edinburgh Scotland, 25-29 Aug. 2014, pp. 732-737.
[29] J. Shan and S. Akella, "3D human action segmentation and recognition using pose kinetic energy", *IEEE Workshop on Advanced Robotics and its Social Impacts*, Evanston, IL, USA, 11-13 Sept. 2014, pp. 69-75.
[30] S. Gaglio, G. L. Re and M. Morana, "Human activity recognition process using 3-D posture data", *IEEE Trans. Human-Machine Systems*, vol. 45, no. 5, pp. 586-597, Oct. 2015, 10.1109/THMS.2014.2377111.
[31] G. I. Parisi, C. Weber and S. Wermter, "Self-organizing neural integration of pose-motion features for human action recognition", *Frontier in Neurobotics*, vol. 9, no. 3, Jun. 2015, 10.3389/fnbot.2015.00003.
[32] E. Cippitelli *et al.*, "A human activity recognition system using skeleton data from RGBD sensors", *Computational Intelligence and Neuroscience*, vol. 2016, p. 1-14, Feb. 2016, 10.1155/2016/4351435.

Cho N. Phyo received a B.C.Sc in Computer Science in 2010 from the University of Computer Studies, Yangon, Myanmar, and from the same university, a B.C.Sc (Hons.) and a M.C.Sc in 2011 and 2014, respectively. Now she is pursuing her Ph.D. in the Interdisciplinary Graduate School of Agriculture and Engineering, University of Miyazaki, Miyazaki, Japan. Her research interests include applications in analyzing human behavior, computer vision, and real-time video surveillance. She is a student member of IEEE, ACM and IAENG.

Thi T. Zin received a B.Sc. degree (with honors) in Mathematics in 1995 from Yangon University, Myanmar, and a M.I.Sc degree in Computational Mathematics in 1999 from the University of Computer Studies, Yangon, Myanmar. She received her Master and Ph.D. degrees in Information Engineering from Osaka City University, Osaka, Japan, in 2004 and 2007, respectively. From 2007 to 2009, she was a Postdoctoral Research Fellow in the Japan Society for the Promotion of Science (JSPS). She is currently a Professor in the Graduate School of Engineering, University of Miyazaki, Miyazaki, Japan. Her research interests include understanding human behavior, ITS, and image recognition. She is a member of IEEE.

Pyke Tin received a B.Sc. degree (with honors) in Mathematics in 1965 from the University of Mandalay, Myanmar, a M.Sc. degree in Computational Mathematics in 1970 from the University of Rangoon, Myanmar and a Ph.D. degree in stochastic processes and their applications in 1976 from Monash University, Australia. He was Rector in the University of Computer Studies, Yangon, and a Professor of Computational Mathematics. He is now a visiting professor in the International Relations Center, University of Miyazaki, Miyazaki, Japan. His research interests include image search engines, queuing systems, computer vision, stochastic processes and their applications to image processing.