Received: 24 October 2018 | Revised: 12 April 2019 | Accepted: 29 July 2019

DOI: 10.4218/etrij.2018-0577

ORIGINAL ARTICLE

ETRI Journal WILEY

Human activity recognition with analysis of angles between skeletal joints using a RGB-depth sensor

Ömer Faruk İnce[1] | Ibrahim Furkan Ince[2] | Mustafa Eren Yıldırım[2,3] | Jang Sik Park[2] | Jong Kwan Song[2] | Byung Woo Yoon[2]

[1]Center for Intelligent and Interactive Robotics, Korea Institute of Science and Technology, Seoul, Rep. of Korea

[2]Department of Electronics Engineering, Kyungsung University, Pusan, Rep. of Korea

[3]Department of Electrical and Electronics Engineering, Bahçeşehir University, Istanbul, Turkey

Correspondence
Jang Sik Park, Department of Electronics Engineering, Kyungsung University, Pusan, Rep. of Korea.
Email: jsipark@ks.ac.kr

Funding information
Busan Brain 21+; Development 3D Map Video Surveillance System for Security Monitoring and Safety Management of Infrastructure Facilities, Grant/Award Number: 2018-0-00488; Institute for Information & Communications Technology Promotion (IITP)

Human activity recognition (HAR) has become effective as a computer vision tool for video surveillance systems. In this paper, a novel biometric system that can detect human activities in 3D space is proposed. In order to implement HAR, joint angles obtained using an RGB-depth sensor are used as features. Because HAR is operated in the time domain, angle information is stored using the sliding kernel method. Haar-wavelet transform (HWT) is applied to preserve the information of the features before reducing the data dimension. Dimension reduction using an averaging algorithm is also applied to decrease the computational cost, which provides faster performance while maintaining high accuracy. Before the classification, a proposed thresholding method with inverse HWT is conducted to extract the final feature set. Finally, the K-nearest neighbor (*k*-NN) algorithm is used to recognize the activity with respect to the given data. The method compares favorably with the results using other machine learning algorithms.

KEYWORDS
activity recognition, dimension reduction, Haar-wavelet transform, K-nearest neighbour, RGB-D sensor

1 | INTRODUCTION

Monitoring and recognition of human activity patterns collected by motion sensors is currently a popular research topic. Human activity recognition (HAR) has been used in several different domains such as robotics [1–3], computer engineering [4,5], healthcare [6], natural sciences [7], and industrial applications [8,9]. Understanding of human activity involves activity recognition and activity pattern discovery. HAR aims to provide highly accurate detection of human activities by adopting a predefined activity model. To do so, a high level conceptual model must first be built and run by structuring an appropriately pervasive system. In contrast, discovery of activity patterns is more closely related to identifying unknown patterns directly from low-level sensor data without applying any predefined models or assumptions.

Previous research has proven that machine learning methodologies work efficiently to classify different activities from sensor data [10–12]. Some types of sensors used in HAR systems are digital cameras, depth sensors, wearable sensors, and gyro sensors [13–19]. Sensor-based systems require two main steps. The first step is the calculation of relevant features based on a sensor's acquired data. In the second step, a chosen classification algorithm defines the activity in accordance with the features obtained in the first step. Common features contain statistics extracted from time-domain signal

analysis, frequency-domain analysis, and wavelet analysis, which is also known as time-frequency analysis.

Vision and other signal-based sensors [20–24] are used to merge technologies with advanced worldwide practical applications and play major roles in science and industry. These technologies provide comfort in daily activities and enhance the quality of life.

This paper proposes a biometric system that uses angles between skeletal joints to recognize human activities in 3D space based on RGB-depth sensor data, which could be useful for elderly care and video surveillance systems. Simply stated, the system obtains particle joint angle pairs and stores them using the sliding kernel method. After creating a feature set, Haar coefficients of the features are obtained, and the number of features is then reduced by applying an averaging dimension reduction technique.

A proposed thresholding method using the inverse Haar wavelet transform (HWT) is applied to enhance the signals for better classification. HWT is commonly applied to filter out data noise, reduce data size, and detect singularities. Thus, HWT is highly effective in time series data processing. Finally, the *k*-nearest neighbor (*k*-NN) algorithm is applied to classify the real-time data originating from the sensor. This paper is organized as follows; Section 1 provides a brief introduction of the proposed study. In Section 2, HAR applications regarding previous approaches are reviewed. In Section 3, the proposed algorithm for human activity recognition is discussed in detail. In Section 4, the experimental environment is presented. In the final section, the advantages and limitations of the proposed study and further improvements for better performance are discussed.

2 | PREVIOUS APPROACHES USING HAR

Human activity recognition has been intensively studied over the past several years. In [25], re-identification of persons in two different mediums was accomplished using SIFT and Bag-of-Features. The use of skin joint features obtained from RGB-depth sensors using HAR to track humans was proposed in [26]. The authors in [27,28] used local and hybrid features for facial analysis. In [29], motion detection was achieved in real-time using multiple cameras.

The majority of HAR studies can be categorized under two types: vision-based and sensor-based methods. A vision-based study is presented in [30] in which the authors estimate the simultaneous pose and shape of articulated objects by using a single depth camera. First, 3D transformations of each skeletal joint are illustrated using a twist map and exponential maps, and the articulated deformation model constructed from the maps is then combined with a probabilistic model to carry out pose tracking. In [31], the authors suggested using the representation of intermediate body parts to map the estimation of poses into a per-pixel classification problem and then restructure the resulting body parts to construct confidence-scored 3D proposals of multiple body joints. Human pose estimation based on a single depth camera was proposed in [32], and relies on the correlations among articulated and generalized Gaussian kernels. The approach consists of embedding the kinematic skeleton into the Gaussian kernels and constructing tree-structured templates from several multivariate Gaussian kernels with quaternion-based rotation.

In [33] a new skeleton-based method is proposed to describe the spatio-temporal aspects of an activity data sequence via the Minkowski and cosine distances between 3D skeletal joints. In [34], multifeatures along with a hidden Markov model (HMM) are used with a single camera for a healthcare application. Spatial-temporal features for HAR were evaluated in [35,36]. In [37], graph formulation is employed for abnormal activity recognition. Although there are some RGB-based studies in the literature, the applications suffer in environments that are totally dark or where illumination changes are present, despite the use of a multi-camera system consisting of eight cameras installed to view a room from every possible angle and to overcome an issue with subject occlusion, for example [33]. However, unlike RGB-based methods, depth-based methods are invariant to illumination changes. In sensor-based studies, authors have proposed the use of multiple accelerometers, wearable sensors, and other types of sensors. In [38], the authors invented a system which is based on systematic performance analysis of motion-sensor-captured behavior for human activity recognition via smart phones. Sensory data sequences using smart phones were collected while participants in the experiment performed typical and daily activities. An activity unit was then characterized by time, frequency, and wavelet domain features. By means of various classification algorithms, both personalized and generalized models were created and performed activity recognition. Sikder and others proposed a new distance metric called the log-sum distance to calculate the difference between two sequences of positive numbers [39]. Basically, the log-sum distance measures the motion data gathered from daily activities. Wearable sensors were used for human behavior analysis in [40]. Another study used pyroelectric sensors for abnormal activity detection [41].

In contrast, internet of things (IoT)-based HAR systems have also been proposed. Subasi and others proposed an intelligent m-healthcare system using IoT technology [42]. The motivation was to provide pervasive human activity recognition using data mining algorithms. Additionally, deep learning algorithms have been recently applied to vision-based HAR systems by Wang, Simonyan and Zisserman, and Karpathy and others in 2014 [43–45]. The authors turn the task of action recognition into one of image classification to handle over-fitting problems caused by the small number of annotated training examples. Because some actions

are highly associated with certain objects and static poses, the static appearance by itself could be a useful solution. Recently, long-short term memory (LSTM) networks have been used in different approaches. Inspired by the success of the weightlessness feature, Tao and others developed two-directional features for bidirectional long short-term memory (BLSTM) to augment learning in HAR systems. For higher accuracy, a new classifier named the multi-column BLSTM (MBLSTM) that combines various acceleration signal features for activity recognition is presented [46].

3 | PROPOSED HUMAN ACTIVITY RECOGNITION MODEL

3.1 | Problem definition

In this study, the main motivation is to develop a HAR system which has low cost and high efficiency. To do so, RGB-depth based system (a camera-based HAR system) using skeletal angle information is proposed. The proposed approach obtains eight different angles of skeletal joints. To obtain the angle information, skeleton detection and tracking is first performed using Microsoft's Kinect SDK 2.0. Because each joint is defined in the SDK, 3D angles can be calculated. Joint pairs are selected with respect to the relevance of basic human activities as defined in related studies. Previous approaches [47,48] show that there is a direct relevance between angle motion sequences and human activities. In particular, the main motivation for using angles is that angles are scale and rotation invariant. The flowchart of the proposed method is shown in Figure 1.

3.2 | Flow analysis of the proposed algorithm

3.2.1 | Angles between joints

A Kinect v2 sensor can detect and track 25 skeletal joints. After the skeleton is detected and tracked, the key part of the method lies in selecting features to create feature vectors. Ofli and others have shown that during standing actions and walking, the right and left knees and the elbow are the most informative features of the human skeleton [49]. In addition, Uddin and others have shown that angles between shoulders, elbows, knees, and the crotch provide useful informative features for 3D human activity recognition [47]. Moreover, Thang and others have used angle pairs of shoulders, elbows, and knees. In the proposed approach, angle pairs of hipbones are also added. The selected angle pairs used in this study are shown in Figure 2.

The value of the angle between two different joints j_1 and j_2 can be calculated by distinguishing the locations of the two joints with respect to a reference joint r in 3D space. The formula is denoted as follows:

$$a_{j_1,j_2} = \cos^{-1}\left(\frac{\vec{rj_1}\cdot\vec{rj_2}}{||rj_1||\cdot||rj_2||}\right). \quad (1)$$

In this equation, $\vec{rj_1}$ represents the distance between joint j_1 and the reference joint r in 3D space. Similarly, $\vec{rj_2}$ represents the distance between joint j_2 and the reference joint r in 3D space. The dot between the vectors indicates the dot product. Lastly, $||rj_1||$ and $||rj_2||$ represent the lengths of these vectors, respectively.

3.2.2 | Sliding Kernel

Because human activity occurs in the time domain, it can be considered as a two dimensional problem. It is thus necessary to observe the angle patterns of each joint angle in the

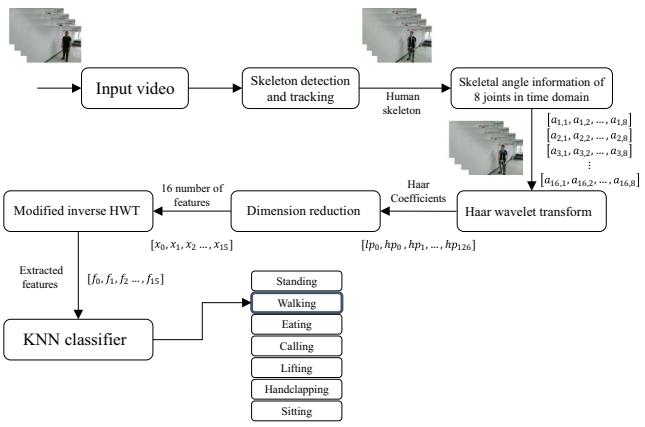

FIGURE 1 Flowchart of the proposed method

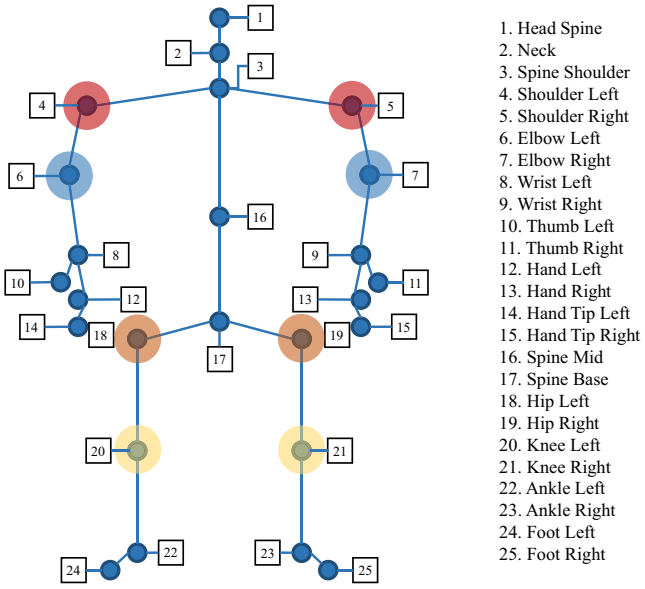

FIGURE 2 Human skeleton and selected angle pairs

time domain to recognize an action. Therefore, storing angle swings in the time domain using the sliding kernel method is beneficial. Angle data of each joint in the time domain can be denoted as follows:

$$a_{f,j} \in N^*, x_{f,j} \leq 360, \tag{2}$$

$$\text{where } f = 1, \ldots, \infty \text{ and } j = 1, 2, \ldots, 8, \tag{3}$$

$$k = \begin{bmatrix} a_{1,1} & \cdots & a_{1,8} \\ \vdots & \ddots & \vdots \\ a_{16,1} & \cdots & a_{16,8} \end{bmatrix} \rightarrow k' = \left[a_{1,1}, a_{1,2}, \ldots, a_{16,8}\right] \rightarrow \left[x_0, x_1, \ldots, x_{127}\right], \tag{4}$$

where $a_{f,j}$ is the angle value of a joint in the jth frame. The number of features obtained in each frame is equal to $8 \times n$, where n represents the kernel size.

In this paper, the sliding kernel is achieved using a queue structure. In each frame, eight joint angles are obtained by the Kinect camera and they are stored in the queue structure. Here the capacity of the queue structure is another parameter of the system, which denotes the kernel size. If the queue capacity is 16, then 16 × 8 = 128 features are stored in the queue. Because the queue structure is a first-in first-out (FIFO) data structure, it behaves as a sliding kernel in the time domain. In our study, the kernel size (in other words the queue capacity) is set to 16, which yields optimum results with respect to time complexity and performance.

3.2.3 | Discrete Haar Wavelet Transform (DWT)

Haar wavelets are square-shaped functions developed to organize data using frequency. Data are transformed from the spatial domain to the frequency domain, and each component on the corresponding resolution scale is stored by wavelet transformation. The basic Haar function $\psi(t)$, with scaling function $\phi(t)$, is defined using the following relationships:

$$\psi(t) = \begin{cases} 1 & t \in [0,0.5), \\ -1 & t \in [0.5,1), \\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

$$\psi_i^j(t) = \sqrt{2^j}\psi(2^j t - i), \tag{6}$$

$$j = 0, 1, \ldots, \infty \text{ and } i = 0, 1, \ldots, 2^j - 1, \tag{7}$$

$$\phi(t) = \begin{cases} 1 & t \in [0,0.5), \\ -1 & t \in [0.5,1), \\ 0 & \text{otherwise}. \end{cases} \tag{8}$$

The functional relevance between the wavelet function and scaling function can be described as follows:

$$\phi(t) = \phi(t) + \phi(2t - 1) \tag{9}$$

and

$$\psi(t) = \psi(t) - \psi(2t - 1). \tag{10}$$

Haar wavelet transform basically consists of averaging and differencing. For an input dataset with 2^n elements, the HWT simply takes the average of each pair of components of the dataset and places them in the first half of a new string (low-pass band). The average differences between each pair of components compose the other half of the new string (high-pass band). This process repeats itself until there are 2^{n-1} detail coefficients and one final sum consisting of low-pass values. Assume that x is a vector with length N, and $x = (x_1, x_2, \ldots, x_N)$, where N is a power of 2. The number of steps required is:

$$\rho = \log_2 N. \tag{11}$$

The calculation of low-pass band coefficients lp_k and high-pass band coefficients hp_k proceeds as follows:

$$lp_k = \frac{\left(x_{2k} + x_{2k+1}\right)}{2} \text{ for } \quad k = 0, 1, \ldots, \frac{N}{2} - 1 \tag{12}$$

and

$$hp_k = \frac{\left(x_{2k} - x_{2k+1}\right)}{2} \text{ for } \quad k = 0, 1, \ldots, \frac{N}{2} - 1. \tag{13}$$

This procedure is also reconstructive because:

$$\frac{\left(x_{2k} + x_{2k+1}\right)}{2} + \frac{\left(x_{2k} - x_{2k+1}\right)}{2} = x_{2k} \tag{14}$$

and

$$\frac{\left(x_{2k} + x_{2k+1}\right)}{2} - \frac{\left(x_{2k} - x_{2k+1}\right)}{2} = x_{2k-1}. \tag{15}$$

In this paper, the number of elements for each input signal is 128, which means that the step number is $\log_2 128 = 7$. After determining the Haar coefficients of the feature vectors, averaging-based dimension reduction is applied.

3.2.4 | Dimension reduction with averaging

After 128 Haar coefficients are produced, the averaging-based dimension reduction method is conducted to decrease the computational cost and eliminate redundancy. The average of two likely HWT candidates in the low-pass module or in the high-pass module will correspond to new elements in the new set of vectors. For instance, the average of the first and second elements will correspond to the first index of the new vector, the average of the third and fourth elements will correspond to the second index of the new vector, and so on. The equation for reducing the dimensions is denoted as follows:

$$f_n = \frac{\left(x_{n,2k} + x_{n,2k+1}\right)}{2}, \text{for} \quad k = 0, 1, \ldots, \frac{N}{2} - 1 \tag{16}$$

and

$$n = 1, 2, \ldots, \rho, \tag{17}$$

where **f** is the feature vector for each reduction. The first subscript of x represents the number of dimension steps, and the second subscript is the index number of a feature vector. Additionally, ρ represents the step number.

In this paper, we use averaging for dimension reduction in the frequency domain. After we obtain the Haar coefficients, we apply the averaging method to reduce the size of the frequency graph of the Haar coefficients. After dimension reduction in the frequency domain, we apply the inverse transform to obtain feature coefficients in the time domain. This is necessary to reduce the dimensions of the feature vectors with minimal loss of information.

3.2.5 | Proposed thresholding with inverse HWT

As mentioned in the Haar Wavelet Transform section, the original signal can be easily reconstructed by applying the reverse averaging and differencing operation without losing information. However, it is possible to obtain a higher degree of compression by setting a non-negative threshold value λ. This method is called lossy compression. Using a threshold will make any low-pass band elements correspond to zero if their magnitude in the transformed signal is less than or equal to the threshold value. By this means, the number of zeros in the transformed signal will increase, which will provide a high level of compression. For lossless compression, the threshold value λ is set to zero. If lossy compression is required or is advantageous, then approximations of the original signal are constructed. Setting the value of λ requires care, as there is a tradeoff between the threshold value and compressed signal quality. The conventional threshold equation is

$$x \in [1, N-1] \tag{18}$$

and

$$T(\lambda, x) = \begin{cases} 0 & \text{if } |x| < \lambda, \\ x & \text{otherwise,} \end{cases} \tag{19}$$

where x represents any Haar coefficient located in the high-pass band.

The proposed thresholding method provides a useful technique to enhance the signal quality while performing inverse HWT. The formulation is

$$x \in [1, N-1], \tag{20}$$

$$\mu = (lp_0 + hp_0 + hp_1 + \ldots + hp_{N-2})/N, \tag{21}$$

$$T(\lambda, x) = \begin{cases} x & \text{if } x > \mu, \\ \text{sign}(x)\mu & \text{otherwise,} \end{cases} \tag{22}$$

where μ is the mean of the Haar coefficients in both the low-pass and high-pass bands.

3.2.6 | Transformation of feature vectors

First, feature vector values contain the information of eight different joint angles. These values are stored (for 16 frames) until 128 features are obtained. To retain the important information stored, data are evaluated in the frequency domain. For time series evaluation, HWT is widely used in signal processing, because HWT compresses the data without losing important information (red signals in Figure 3). In the frequency domain, Haar coefficients are averaged for dimension reduction (green signals in Figure 3). Thus, the important information is kept and redundancy will be eliminated. In other words, the power spectrum of each signal will be reduced, and this provides a low dynamic range evaluation of Haar coefficients. After dimension reduction, a novel thresholding method is proposed. In this method, the Haar coefficients below the mean of all Haar coefficients are replaced with the mean of all Haar coefficients, but the sign of each replaced coefficient is kept. Through the inverse HWT, the feature extraction process is performed (purple signals in Figure 3). In summary, a new feature vector contains the characteristics of the raw data, but at a different scale and shape. In Figure 3, purple signals represent the results of the thresholding methods combined with inverse HWT. Even though both methods eliminate redundancy and provide low dynamic range evaluation, it is clear that the proposed method creates more peaks and valleys in the data, which indicate more efficient classification performance.

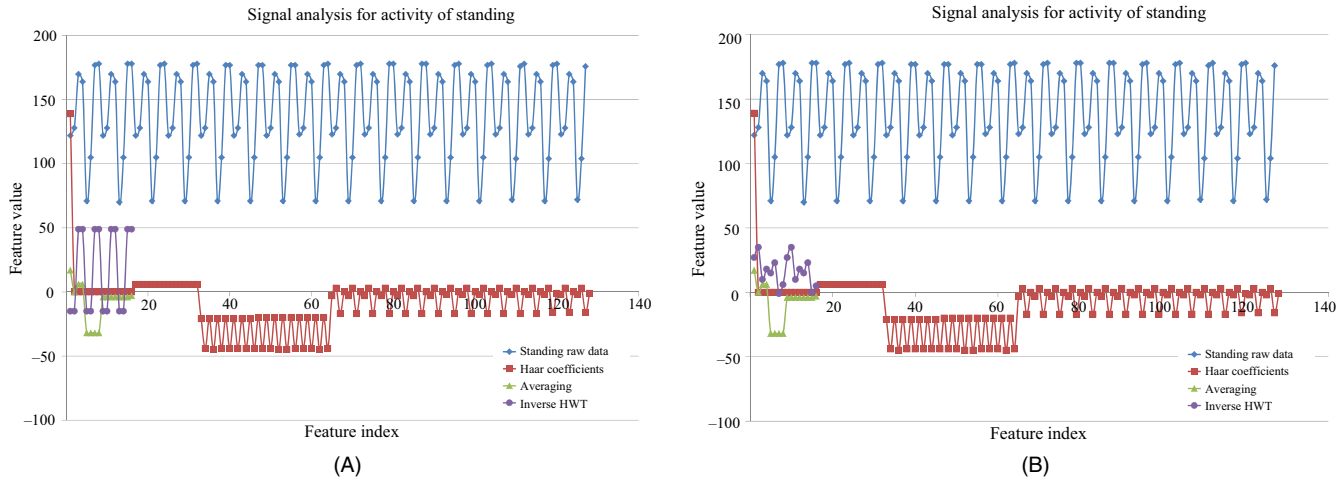

FIGURE 3 Thresholding comparison: (A) conventional method and (B) proposed method

4 | EXPERIMENTAL RESULTS AND CONSIDERATIONS

4.1 | Experimental environment and dataset

The dataset for this study was created by Kyungsung University, Department of Electronic Engineering. For implementation of the proposed HAR system, Microsoft Visual Studio 2013, C# with Microsoft Kinect SDK 2.0, Vitruvius, and the Accord.net machine learning framework are used. For comparisons with different parameters of *k*-NN and other algorithms such as Random Forest (RF) and genetic algorithm (GA), Weka data mining software is used. Lastly, Python 3.6.4 with Tensorflow is also used for the LSTM network performance evaluation. This dataset contains information regarding 10 different people whose height, weight, and clothing are different. Each activity is recorded for each person (user) two times to construct the training set and one time to construct the test set. The reason is that test sequences include different angle variations of the same activities, which are used to evaluate the accuracy of the proposed system in real-world conditions. Moreover, for *k*-fold cross validation, training and test sequences were combined, and k was set to 10. Because each person performs an activity at a different pace, the number of instances for each activity is different. The number of instances for each activity used in the dataset is shown in Table 1.

For the preparation of the training and test sets, the Microsoft Kinect v2 sensor is placed at a height of 1.70 m. During the creation of the training set, a user is recorded in two directions, which are ±45° with respect to the camera. However, the users are recorded for the entire area between ±45° and −45° while preparing the test set. The key motivation behind the use of test activity patterns not used for training is that the proposed method provides rotation invariance. The dataset setup is presented in Figure 4.

TABLE 1 Number of instances used in the dataset for each activity

	Train	Test
Activity	**Number of instances**	**Number of instances**
Standing	2018	848
Walking	2060	955
Eating	1728	926
Calling	1692	873
Lifting	2264	832
Handclapping	2224	859
Sitting	2183	969
Total	14 169	6262

Finally, the dataset is composed of two activity primitives, which are posture and motion. The list of primitives is given in Table 2 and images of each activity in the dataset are shown in Figure 5.

4.2 | Comparison of different *k*-NN parameters

As mentioned previously, the *k*-NN algorithm is used as the classifier in this paper. The main reason for choosing *k*-NN as the classifier is that *k*-NN is a lazy learner. This means that the *k*-NN algorithm does not model the data while training. However, it remembers the data during the test phase. Another reason is that *k*-NN is quite effective in a large number of applications; however, as the number of dimensions increases, *k*-NN performance deteriorates.

Because there are various parameters associated with *k*-NN, a short explanation of which parameters are used for

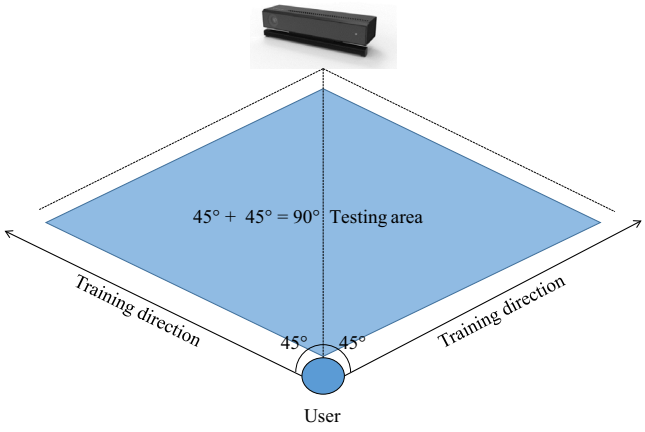

FIGURE 4 Dataset setup for training and testing

TABLE 2 Activity primitives and their members

Posture	Motion
Standing	Walking
Calling	Eating
Sitting	Lifting
—	Handclapping

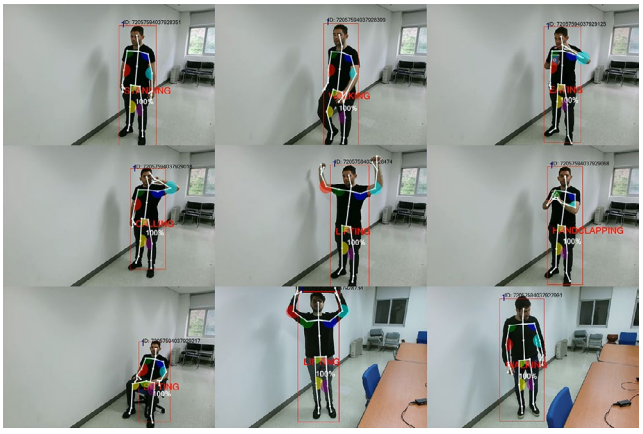

FIGURE 5 HAR-activity examples from dataset

this study follows. First, because *k*-NN looks for the nearest neighbor in a cluster by calculating the distance between attributes, one of the main parameters is distance. There are various types of distance metrics, such as the Euclidean distance, Manhattan distance, or Minkowski distance. Each of these metrics has advantages and disadvantages, and selecting which to use largely depends on the dataset. In this study, the Euclidean distance was selected based on the analysis results discussed below.

After choosing the distance metric, assigning the value of *k* (the number of closest training examples to be considered as similar to a test sample) is the next task, and assigning the value of *k* also depends on the dataset. There are some ways to find the optimum *k* value, such as cross-validation and resampling. Assigning *k* is a critical task. A small k value will allow noise to have a higher influence on the data. In contrast, a large *k* value increases computational complexity and obviates the main philosophy behind *k*-NN (attributes which are close to each other could have similar densities or classes). In the proposed study, the k value assigned is 1. To determine the best performance, the data were tested with different *k* values and different distance metrics. Performance results from this comparison are shown in Table 3.

4.3 | Comparison of different classifiers

In this section, the comparison of different classifiers is presented. For this evaluation, *k*-NN, RF, and LSTM networks are tested on the same dataset. The motivation for comparing *k*-NN and RF is that there is a close relationship between these two algorithms [50]. Simply stated, both algorithms can be viewed as weighted neighborhood schemes. They create models from training data to make predictions for new observations by checking the neighborhoods, and formalize a weight function. RF also provides a variable ranking mechanism that can be used to select important variables. Because *k*-NN is a nonparametric model, it is usually a good classifier for many situations in which the joint distribution is unknown or difficult to model parametrically. This is especially true for high-dimension datasets.

However, there is some dissimilarity as well. For instance, RF produces an in-memory classification model which does not require database lookups, while *k*-NN uses on-the-spot learning that requires extensive computations, which makes *k*-NN inefficient for classifying large databases. Apart from this, eager learners such as RF cannot easily model decision spaces with complicated decision boundaries; in contrast, *k*-NN performs instance-based learning which leads to accurate performance if a well-tuned *k*-NN model is used. Additionally, *k*-NN is a lazy learning scheme, and only stores the input data during the training process; thereby, training is essentially spontaneous [51]. The parameters used for *k*-NN were given in the previous section. For RF, the size of each bag was set to 100, and 100 trees were used. Additionally, the number of folds for back fitting was set to 0 (no back fitting). In addition to these two algorithms, LSTM was included in this comparison. The reason is that deep learning algorithms have become very popular recently in the machine learning field, and they usually achieve higher accuracy rates compared to other classifiers. In this study, a many-to-one recurrent neural net (RNN) architecture with two LSTM cells was used. The number of hidden layers was set to 32, and the learning rate and lambda loss amount were set to 0.0015 and

TABLE 3 Performance using different *k* values and metrics

Metric	*k* value	Accuracy (%)	F-1 (%)	Precision (%)	Recall (%)
Euclidean distance	1	86.1	85.7	86.0	86.1
	3	85.6	85.1	85.4	85.6
	5	84.9	84.3	84.8	84.9
	9	84.2	83.6	84.0	84.2
Manhattan distance	1	85.9	85.6	85.8	85.9
	3	85.6	85.1	85.4	85.6
	5	84.8	84.2	84.6	84.8
	9	84.1	83.6	84.0	84.2
Minkowski distance	1	86.1	85.7	86.0	86.1
	3	85.6	85.1	85.4	85.6
	5	84.9	84.3	84.8	84.9
	9	84.2	83.6	84.0	84.2

0.0025, respectively. A comparison was also made between the feature vectors obtained by; (i) proposed method with proposed thresholding method, (ii) proposed method with conventional thresholding method, (iii) dimension reduced data with averaging method, and (iv) dimension reduced data with GA. The performances of these feature vectors using the *k*-NN classifier are shown in Figure 6 based on confusion matrices. Briefly, a confusion matrix evaluates the quality of the predictions of a classifier on a given dataset. The diagonal units mean the number of points for which the estimated labels represent true positives, while the off-diagonal elements are mislabeled by the classifier. Thus, the higher the diagonal values of the confusion matrix are, the better the accuracy. The comparison results of the three different classifiers with the four different feature vectors are presented in Table 4. As seen in Table 4, *k*-NN achieves the best accuracy for this problem, and because it is a lazy learner and does not model the data during training, the speed of the algorithm is high. The accuracies of RF and LSTM are very similar for this dataset. Cross-validation experiments were also conducted with the same parameters, and the results are listed in Table 5.

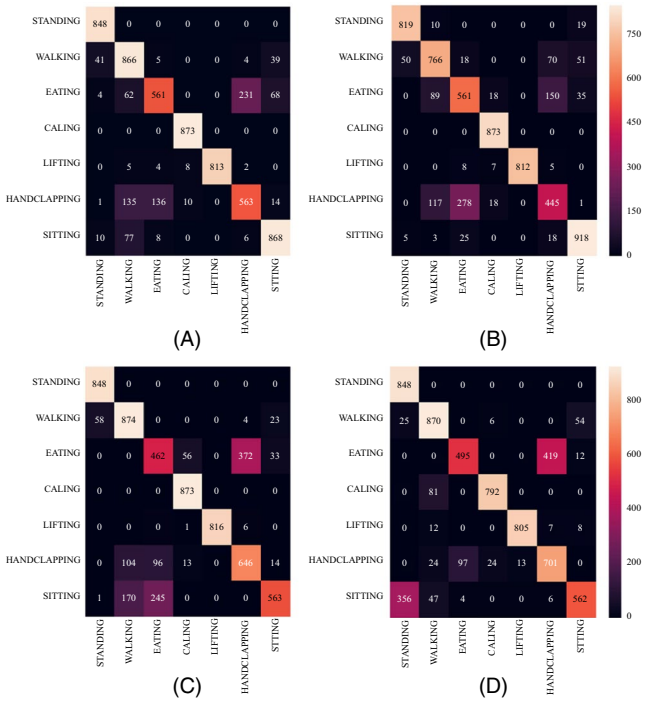

FIGURE 6 Comparison of feature vectors generated by (A) proposed method with proposed thresholding method; (B) proposed method with conventional thresholding method; (C) dimension reduced data with averaging method; and (D) dimension reduced data with GA

LSTM was expected to achieve higher accuracy, but did not. The reason could be that neural networks are designed for high-dimension datasets. Because the dataset used here has low dimension (seven classes), LSTM may have difficulties in constructing a proper model. In contrast, *k*-NN can suffer from the curse of dimensionality. This means that although *k*-NN is effective in a large number of cases, its accuracy decreases as the dimension (number of classes) increases. It is thus likely that the low number of classes used to represent the dataset allowed *k*-NN to be slightly more accurate than LSTM.

In contrast, as a feature vector, the proposed thresholding provides better results than conventional thresholding in all comparisons. Moreover, similar studies presented in the literature have produced various levels of performance. A Gaussian mixture-based HMM for human daily activity recognition study [52] obtained 84% recall accuracy, whereas using a depth video sensor for indoor activity recognition achieved 90.33% accuracy [53]. In [54], the authors used spatiotemporal multi-fused features from depth video and obtained a 94.1% accuracy rate. Other depth-based studies [55–57] reported recognition rates of 91.29%, 78.5%, and 83.9%. In [58], an 89.1% recognition rate was achieved by using a combined support vector machine (SVM) and HMM architecture.

TABLE 4 Results of different classifiers with different feature vectors

Classifier	Feature vector	Accuracy (%)	F-1 (%)	Precision (%)	Recall (%)
k-NN $k = 1$	Proposed method with proposed thresholding	86.1	85.7	86.0	86.1
	Proposed method with conventional thresholding	84.1	83.7	83.6	84.1
	Dimension reduced data with averaging	81.0	80.4	81.6	81.0
	Dimension reduced data with GA	81.0	80.5	83.2	81.0
Random Forest	Proposed method with proposed thresholding	82.6	82.4	82.4	82.6
	Proposed method with conventional thresholding	80.6	80.3	80.3	80.6
	Dimension reduced data with averaging	80.8	80.3	81.4	80.8
	Dimension reduced data with GA	82.9	82.2	85.6	82.9
LSTM	Proposed method with proposed thresholding	82.2	81.9	82.6	82.2
	Proposed method with conventional thresholding	78.7	78.1	78.1	78.7
	Dimension reduced data with averaging	45.5	42.7	50.4	45.5
	Dimension reduced data with GA	76.3	75.6	77.8	76.3

TABLE 5 Cross-validation results of different classifiers with different feature vectors

Classifier	Feature Vector	Accuracy (%)	F-1 (%)	Precision (%)	Recall (%)
k-NN $k = 1$	Proposed method with proposed thresholding	97.8	97.8	97.8	97.8
	Proposed method with conventional thresholding	93.2	93.2	93.3	93.2
	Dimension reduced data with averaging	86.6	86.5	86.7	86.6
	Dimension reduced data with GA	99.5	99.5	99.5	99.5
Random Forest	Proposed method with proposed thresholding	98.1	98.1	98.1	98.1
	Proposed method with conventional thresholding	93.9	94.0	94.0	93.9
	Dimension reduced data with averaging	89.7	89.7	89.7	89.7
	Dimension reduced data with GA	99.9	99.9	99.9	99.9
LSTM	Proposed method with proposed thresholding	99.1	99.2	99.2	90.8
	Proposed method with conventional thresholding	94.0	94.1	94.1	94.0
	Dimension reduced data with averaging	68.2	72.1	76.6	68.2
	Dimension reduced data with GA	99.9	99.9	99.9	99.9

5 | CONCLUSIONS

Recognition of human activity has become one of the most popular research topics in the machine learning field. Generally, proposed approaches focus on heterogeneous and/or large scale data. The creation of a HAR system is a nontrivial problem based on rotation and scale variations, complex camera motion, large interclass variations, and data margin issues.

In this study, a new method is proposed for human activity recognition using angle patterns between skeletal joints. The reason behind this approach is that angles are scale and rotation invariant features. Because human activity occurs in the time domain, these angle values are stored using the sliding kernel method. Stored kernel elements are then evaluated in the frequency domain. For time series evaluation, we applied HWT to compress the data without losing information. By conversion to the frequency domain, the number of Haar coefficients was reduced from 128 to 16 to lower the computational cost. Using this method, the important information is retained and redundancy is eliminated. After dimension reduction, a novel thresholding method for feature extraction is proposed. Feature extraction is accomplished through the inverse HWT. Then, the k-NN algorithm is used to recognize human activities. Various classifiers and different feature vectors are compared as a cross-check to evaluate whether the proposed method is a good choice for human activity recognition. According to the experimental results, the best accuracy achieved by the proposed method was 86.1%. In summary, the various steps necessary to construct HAR systems have been reviewed. The proposed automatic human activity recognition system using a RGB-depth camera appears to be suitable for video surveillance systems and in elderly care. A major contribution of the proposed method involves

reducing the data dimension in the frequency domain so that analysis can be conducted in a low dynamic range. Extensive testing results have verified that the proposed method performed adequately.

However, there are some disadvantages associated with the proposed algorithm. First, incorrect skeleton detection causes incorrect angle calculations and thus automatically misleads the classification. Because the system is trained based on activities in two directions at different angles and positions, some confusion can occur when attempting to recognize an activity because of possible similarities in the angles and positions of different activities. Finally, considering price and comfort, a simple RGB camera rather than a RGB-depth camera could be considered.

For future work, evaluating the performance using a simple RGB camera could extend the application areas of the proposed HAR system.

ACKNOWLEDGMENTS

This work was supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea Government (MSIP) (No. 2018-0-00488, Development 3D Map Video Surveillance System for Security Monitoring and Safety Management of Infrastructure Facilities) and Brain Busan 21 + 2019.

ORCID

Mustafa Eren Yıldırım https://orcid.org/0000-0002-0662-2770

Jang Sik Park https://orcid.org/0000-0003-1794-7631

REFERENCES

1. M. T. Uddin and M. A. Uddin, *Human activity recognition from wearable sensors using extremely randomized trees*, in Proc. Int. Conf. Electr. Eng. Inf. Commun. Technol., Dhaka, Bangladesh, May 2015, pp. 769–778.
2. A. Jalal et al., *Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home*, Indoor Built Environ. **22** (2013), no. 1, 271–279.
3. Y. Zhan and T. J. Kuroda, *Wearable sensor-based human activity recognition from environmental background sounds*, J. Ambient Intell. Humanized Comput. **5** (2014), no. 1, 77–89.
4. Z. A. Jalal and I. Uddin, *Security Architecture for Third Generation (3G) using GMHS Cellular Network*, in Proc. Int. Conf. on Emerging Technol., Islamabad, Pakistan, Nov. 2007, pp. 74–79.
5. A. Jalal and M. A. Zeb, *Security Enhancement for E-learning portal*, Int. J. Comput. Sci. Netw. Security **8** (2008), no. 3, 41–45.
6. A. Jalal and M. A. Zeb, *Collaboration achievement along with performance maintenance in video streaming*, in Proc. Int. Conf. Comput. Inf. Technol., Dhaka, Bangladesh, 2007, pp. 369–374.
7. A. Jalal and A. Shahzad, *Multiple facial feature detection using vertex-modeling structure*, in Proc. IEEE Conf. Interactive Comput. Aided Learn., Villach, Austria, Sept. 2007, pp. 26–28.
8. A. Jalal, S. Kim, and B. J. Yun, *Assembled algorithm in the real-time h.263 codec for advanced performance*, in Proc. Int. Workshop Enterprise Netw. Comput. Healthcare Industry, Busan, Rep. of Korea, June 2005, pp. 295–298.
9. A. Jalal and S. Kim, *Algorithmic implementation and efficiency maintenance of real-time environment using low-bitrate wireless communication*, in Proc. IEEE Workshop Softw. Technol. Future Embedded Ubiquitous Syst., Gyeongju, Rep. of Korea, Apr. 2006, pp. 81–88.
10. N. Ravi et al., *Activity recognition from accelerometer data*, in Proc. Conf. Innovative Applicat. Artif. Intell., Pittsburgh, PA, USA, July 2005, pp. 1541–1546.
11. D. Figo et al., *Preprocessing techniques for context recognition from accelerometer data*, Personal Ubiquitous Comput. **14** (2010), 645–662.
12. Ç. B. Erdaş et al., *Integrating features for accelerometer-based activity recognition*, Procedia Comput. Sci. **98** (2016), 522–527.
13. D. Koller et al., *Real-time vision-based camera tracking for augmented reality applications*, in Proc. ACM Symp. Virtual Reality Softw. Technol., Lausanne, Switzerland, Sept. 1997, pp. 87–94.
14. A. Jalal, M. Z. Uddin, and T. Kim, *Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home*, IEEE Trans. Consumer Electron. **58** (2012), no. 3, 863–871.
15. S. Kamal, A. Jalal, and D. Kim, *Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM*, J. Electr. Eng. Technol. **11** (2016), no. 6, 1857–1862.
16. A. Jalal, Y. Kim, and D. Kim, *Ridge body parts features for human pose estimation and recognition from RGB-D video data*, in Proc. Int. Conf. Comput., Commun. Netw. Technol., Hefei, China, July 2014, pp. 1–6.
17. A. Jalal et al., *Human activity recognition via the features of labeled depth body parts*, Lecture Notes Comput. Sci. **7251** (2012), 246–249.
18. A. Jalal, T. K. Jeong, and T. S. Kim, *Development of a life logging system via depth imaging-based human activity recognition for smart homes*, in Proc. Int. Symp. Sustainable Healthy Buildings, Seoul, Rep. of Korea, Sept. 2012, pp. 91–95.
19. A. Jalal and S. Kamal. *Real-time life logging via a depth silhouette-based human activity recognition system for smart home services*, in Proc. Int. Conf. Adv. Video Signal Based Surveillance, Seoul, Rep. of Korea, Aug. 2014, pp. 74–80.
20. J. L. Johnson, *Design of experiments and progressively sequenced regression are combined to achieve minimum data sample size*, Int. J. Hydromechatronics **1** (2018), no. 3, 308–331.
21. L. Alberto, S. Vincentelli, and B. Vigna, *Autonomous vehicles: A playground for sensors*, in Proc. Int. Workshop Adv. Sens. Interfaces, Vieste, Italy, June 2017, p. 2.
22. J. L. Johnson, *Reynolds stress statistics in the near nozzle region of coaxial swirling jets*, Int. J. Hydromechatronics **1** (2018), no. 3, 332–349.
23. V. Lumelsky, *Whole-body robot sensing and human-robot interaction*, in Proc. Int. Symp. Micro-NanoMechatronics Human Sci., Nagoya, Japan, Nov. 2012, pp. 155–155.

24. S. Huang et al., *Wear calculation of sandblasting machine based on EDEM-FLUENT coupling*, Int. J. Hydromechatronics **1** (2018), no. 4, 447–459.
25. Q. Huang, J. Yang, and Y. Qiao, *Person re-identification across multi-camera system based on local descriptors*, in Proc. Int. Conf. Distribut. Smart Cameras, Hong Kong, China, Oct. 2012, pp. 1–6.
26. A. Farooq, A. Jalal, and S. Kamal, *Dense RGB-D map-based human tracking and activity tecognition using skin joints features and self-organizing map*, KSII Trans. Internet Inf. Syst. **9** (2015), no. 5, 1856–1869.
27. A. Jalal and S. Kim, *Global security using human face understanding under vision ubiquitous architecture system*, World Academy Sci. Eng. Technol. **2** (2008), no. 1, 160–164.
28. F. Farooq, J. Ahmed, and L. Zheng, *Facial expression recognition using hybrid features and self-organizing maps*, in Proc. IEEE Int. Conf. Multimedia Expo, Hong Kong, China, July 2017, pp. 409–414.
29. H. Yoshimoto, N. Date, and S. Yonemoto, *Vision-based real-time motion capture system using multiple cameras*, in Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst., Tokyo, Japan, Aug. 2003, pp. 247–251.
30. M. Ye and R. Yang, *Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera*, in Proc. IEEE Conf. Computer Vision Pattern Recogn., Columbus, OH, USA, June 2014, pp. 2345–2352.
31. J. Shotton et al., *Real-time human pose recognition in parts from single depth images, Machine Learning for Computer Vision*, Studies in Computational Intelligence **411** (2013), 119–135.
32. M. Ding and G. Fan, *Articulated and generalized Gaussian kernel correlation for human pose estimation*, IEEE Trans. Image Process. **25** (2016), no. 2, 776–789.
33. Y. Hbali et al., *Skeleton-based human activity recognition for elderly monitoring systems*, IET Comput. Vision **12** (2018), no. 1, 16–26.
34. A. Jalal, S. Kamal, and D. Kim, *A depth video-based human detection and activity recognition using multi-features and embedded hidden Markov models for health care monitoring system*, Int. J. Interactive Multimedia Artif. Intell. **4** (2017), no. 4, 54–62.
35. T. N. Nguyen and N. Q. Ly, *Abnormal activity detection based on dense spatial-temporal features and improved one-class learning*, in Proc. Int. Symp. Inf. Commun. Technol., Nha Trang City, Viet Nam, Dec. 2017, pp. 370–377.
36. D. Singh and C. K. Mohan, *Graph formulation of video activities for abnormal activity recognition*, Pattern Recogn. **65** (2017), 265–272.
37. A. Jalal, M. Maria, and M. Sidduqi, *Robust spatio-temporal features for human interaction recognition via artificial neural network*, in Proc. Int. Conf. Frontiers Inf. Technol., Islamabad, Pakistan, Dec. 17–19, 2018, pp. 218–223.
38. Y. Chen and C. Shen, *Performance analysis of smartphone-sensor behavior for human activity recognition*, IEEE Access **5** (2017), 3095–3110.
39. F. Sikder and D. Sarkar, *Log-sum distance measures and its application to human-activity monitoring and recognition using data from motion sensors*, IEEE Sensors J. **17** (2017), no. 14, 4520–4533.
40. A. Jalal et al., *Wearable sensor-based human behavior understanding and recognition in daily life for smart environments*, in Proc. Int. Conf. Frontiers Inf. Technol., Islamabad, Pakistan, Dec. 17–19, 2018, pp. 105–110.
41. X. Luo et al., *Abnormal activity detection using pyroelectric infrared sensors*, Sensors **16** (2016), 1–17.
42. A. Subasi et al., *IoT based mobile healthcare system for human activity recognition*, in Proc. Learn. Technol. Conf. (L&T), Jeddah, Saudi Arabia, Feb. 2018, pp. 29–34.
43. K. Wang et al., *3D human activity recognition with reconfigurable convolutional neural networks*, in Proc. ACM Int. Conf. Multimedia, Orlando, FL, USA, Nov. 2014, pp. 97–106.
44. K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556. 2014.
45. A. Karpathy et al., *Large-scale video classification with convolutional neural networks*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Columbus, OH, USA, June 23–28, 2014, pp. 1725–1732.
46. D. Tao, Y. Wen, and R. Hong, *Multicolumn bidirectional long short-term memory for mobile devices-based human activity recognition*, IEEE Internet Things J. **3** (2016), no. 6, 1124–1134.
47. N. D. Thang et al., *Estimation of 3-D human body posture via coregistration of 3-D human model and sequential stereo information*, Appl. Intell. **35** (2011), no. 2, 163–177.
48. Md Z Uddin, N. D. Thang, and T.-S. Kim, *Human Activity Recognition via 3-D joint angle features and Hidden Markov models*, in Proc. Int. Conf. Image Process., Hong Kong, China, Sept. 2010, pp. 713–716.
49. F. Ofli et al., *Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition*, J. Visual Commun. Image Representation **25** (2014), no. 1, 24–38.
50. Y. Lin and Y. H. Jeon, Random forests and adaptive nearest neighbors, Technical Report No. 1055, University of Wisconsin, 2002.
51. O. F. Ince et al., *Human identification using video-based analysis of the angle between skeletal joints*, J. Institute Contr. Robot. Syst. **24** (2018), no. 3, 263–270.
52. L. Piyathilaka and S. Kodagoda, *Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features*, in Proc. Conf. Industrial Electron. Applicat., Melbourne, Australia, June 2013, pp. 567–572.
53. A. Jalal, S. Kamal, and D. Kim, *A depth video sensor-based lifelogging human activity recognition system for elderly care in smart indoor environments*, Sensors **14** (2014), no. 7, 11735–11756.
54. A. Jalal et al., *Robust human activity recognition from depth video using spatiotemporal multi-fused features*, Pattern Recogn. **61** (2017), 295–308.
55. A. Jalal, S. Kamal, and D. Kim, *Shape and motion features approach for activity tracking and recognition from kinect video camera*, in Proc. Int. Conf. Adv. Inf. Netw. Applicat. Workshops, Gwangiu, Rep. of Korea, Mar. 2015, pp. 445–450.
56. A. Jalal and Y. Kim, *Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data*, in Proc. Int. Conf. Adv. Video Signal Based Surveillance, Seoul, Rep. of Korea, Aug. 2014, pp. 119–124.
57. A. Jalal, S. Kamal, and D. Kim, *Individual detection-tracking-recognition using depth activity images*, in Proc. Int. Conf. Ubiquitous Robots Ambient Intell., Goyang, Rep. of Korea, Oct. 2015, pp. 450–455.
58. H. Wu et al., *Human activity recognition based on the combined SVM&HMM*, in Proc. Int. Conf. Inf. Auto., Hailar, China, July 2014, pp. 219–224.

AUTHOR BIOGRAPHIES

Ömer Faruk İnce received his BS degree in Electrical and Electronics Engineering from Isik University, Istanbul, Turkey in 2012. In 2015 and 2018, he, respectively, received his MS and PhD degrees in Electronics Engineering from Kyungsung University, Korea. After earning his PhD, he joined and continues to work for the Korea Institute of Science and Technology in the Center for Intelligent & Interactive Robotics as a post-doctoral researcher.

Ibrahim Furkan Ince received his PhD degree in IT convergence design from the Graduate School of Digital Design, Kyungsung University, Pusan, Rep. of Korea, in 2010. For post-doctoral studies, he participated in research activities at the University of Tokyo, Japan, from 2010 to 2012. He worked as a chief research engineer at Hanwul Multimedia Communication Co. Ltd., Pusan, Rep. of Korea, from May 2012 to May 2014. Additionally, he worked as an assistant professor of computer engineering at Gediz University, Izmir, Turkey, from November 2014 to July 2016. Currently, he is working as a freelance researcher for the Department of Electronics, Kyungsung University, Pusan, Rep. of Korea. His research interests include image processing, computer vision, pattern recognition, and human-computer interaction.

Mustafa Eren Yıldırım received his BS degree in Electrical Engineering from Bahcesehir University, Istanbul, Turkey, in 2008 and his MS and PhD degrees in Electronics Engineering from the Graduate School of Electrical and Electronics Engineering, Kyungsung University, Pusan, Rep. of Korea, in 2010 and 2014, respectively. He worked as a researcher and lecturer for Kyungsung University until August 2015. He is currently holding two assistant professor positions in the Department of Electrical and Electronics Engineering, Bahcesehir University and the Department of Electronics Engineering, Kyungsung University. His research interests include image processing, computer vision, and pattern recognition.

Jang Sik Park received his BS, MS, and PhD degrees in Electronic Engineering from Busan National University in 1992, 1994, and 1999, respectively. From 1997 to 2011 he was a professor at Dongeui Institute of Technology. Since 2011, he has been a professor in the Dept. of Electrical and Electronics Engineering, Kyungsung University. His research interests include machine learning, video/image processing and understanding, speech and audio signal processing, and embedded systems.

Jong Kwan Song received his BS degree in Electronics from Busan University in 1989 and his MS degree in Electronics from KAIST in 1991. He received his PhD degree in Electronics from KAIST in 1995. From 1995 to 1997, he was a researcher at Korea Mobile Telecom. Since 1997, he has been a professor in the Dept. of Electrical and Electronics Engineering, Kyungsung Univ. His research interests include video/image processing, nonlinear digital signal processing, and embedded systems.

Byung Woo Yoon received his BS, MS, and PhD degrees in Electronic Engineering from Busan National University, Busan, Rep. of Korea, in 1987, 1989, and 1992, respectively. From 1993 to 1995, he was a senior researcher at the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea, where he was involved in development for the CDMA mobile communication systems group. He was a visiting scholar at the Department of Electrical Engineering, University of Colorado, USA from 2001 to 2002, and at the Department of Electrical Engineering, University of North Carolina, USA, from 2008 to 2009. Since 1995, he has been a professor at Kyungsung University, Busan, Rep. of Korea. His research interests include signal processing, design of VLSI, and development of digital systems.