applied sciences

MDPI

Article

Performance Boosting of Scale and Rotation Invariant Human Activity Recognition (HAR) with LSTM Networks Using Low Dimensional 3D Posture Data in Egocentric Coordinates

Ibrahim Furkan Ince [1,2]

[1] Department of Electronics Engineering, Kyungsung University, Busan 48434, Korea; furkanince@ks.ac.kr
[2] Department of Computer Engineering, Nisantasi University, 34485 Istanbul, Turkey

Received: 3 November 2020; Accepted: 25 November 2020; Published: 27 November 2020

Abstract: Human activity recognition (HAR) has been an active area in computer vision with a broad range of applications, such as education, security surveillance, and healthcare. HAR is a general time series classification problem. LSTMs are widely used for time series classification tasks. However, they work well with high-dimensional feature vectors, which reduce the processing speed of LSTM in real-time applications. Therefore, dimension reduction is required to create low-dimensional feature space. As it is experimented in previous study, LSTM with dimension reduction yielded the worst performance among other classifiers, which are not deep learning methods. Therefore, in this paper, a novel scale and rotation invariant human activity recognition system, which can also work in low dimensional feature space is presented. For this purpose, Kinect depth sensor is employed to obtain skeleton joints. Since angles are used, proposed system is already scale invariant. In order to provide rotation invariance, body relative direction in egocentric coordinates is calculated. The 3D vector between right hip and left hip is used to get the horizontal axis and its cross product with the vertical axis of global coordinate system assumed to be the depth axis of the proposed local coordinate system. Instead of using 3D joint angles, 8 number of limbs and their corresponding 3D angles with X, Y, and Z axes of the proposed coordinate system are compressed with several dimension reduction methods such as averaging filter, Haar wavelet transform (HWT), and discrete cosine transform (DCT) and employed as the feature vector. Finally, extracted features are trained and tested with LSTM (long short-term memory) network, which is an artificial recurrent neural network (RNN) architecture. Experimental and benchmarking results indicate that proposed framework boosts the performance of LSTM by approximately 30% accuracy in low-dimensional feature space.

Keywords: human activity recognition (HAR); Kinect depth sensor; 3D posture data; egocentric coordinate system; dimension reduction; discrete cosine transform (DCT); deep learning; LSTM

1. Introduction

Human activity recognition (HAR) is one of the most essential topics of computer vision concerning the last two decades and has been used in various areas such as video-based surveillance systems [1], elderly care [2], education [3], and healthcare [4–7]. HAR provides information about human physical activity and aims to discover simple or complex procedures in a very realistic environment. To recognize human activities at the highest accuracy, HAR presents the right diagnosis of activity models obtained from various sensors. Sensors used in HAR applications consist of three clusters that are cameras, wearable sensors, and gyro sensors [8–12]. General approaches address a HAR problem in two main categories as vision-based and non-vision based systems. Vision-based HAR systems combine different methods with advanced applications using image processing. However, non-vision based HAR

systems extract the relevant features coming from the sensor and recognize the activity using a proper machine learning classifier. Both methods have positive and negative sides compared to each other. For instance, non-vision based systems work better in terms of environmental conditions such as fixed scenes and lack of lighting and occlusion. On the other hand, vision-based sensors are much cost effective and are more useful in daily life applications (video surveillance systems) [13]. For this study, we decided on using the vision-based sensor since vision-based sensors fit more for daily life use and are more affordable. There have been various studies for vision-based human activity recognition in the literature. Despite the fact that there exist human activity recognition systems based on RGB cameras, researches indicate that dark environment and illumination changes are still challenging problems. In order to overcome this issue, illumination-invariant systems have been developed using depth sensors. In this regard, a human posture recognition system based on a single depth camera was presented in [14] where skeleton information is articulated with rotated Gaussian kernels and indexed in tree structure. Another study conducting the same issue is also presented in [15] where 3D transformations of each skeletal joint are indexed in twist and exponential maps to construct a deformation model to be employed for the recognition and pose tracking of the objects within the range of a single depth camera. In addition, spatiotemporal behavior of human activities is extracted by depth sensors using the cosine distances among 3D skeleton joints [16]. Similarly, in [17], 3D pose estimation system is proposed in which multiple body joints are constructed in a per-pixel classification problem by combining confidence scored intermediate body parts.

Different from RGB cameras and depth sensors, various sensors have been employed to achieve human activity recognition such as multiple accelerometers, pyroelectric sensors, and wearable sensors. In [18], authors use wearable sensors to perform human behavior analysis. In [19], motion sensors are employed to analyze daily motion activities. Another study employed pyroelectric sensors to recognize abnormal human activities [20]. In [21], smart phones are used for human activity recognition based on the analysis of signals that come from motion sensors. Additionally, internet of things (IoT) technology is used for human activity recognition by employing different machine learning methods [7].

On the other hand, various feature types, data structures and machine learning approaches were employed to obtain better performance in human activity recognition. In [6], a healthcare application based on a single camera is proposed in which multiple features are classified by means of a Hidden Markov Model (HMM). In other study, spatial-temporal features are extracted and analyzed in [22,23]. In [24], graph structure is employed for abnormal human activity recognition. Additionally, a system based on weighted segmentation of the red channel is proposed to control background noise in which feature selection is performed by averaging and ranking the correlation coefficients of background and foreground weights [25]. Moreover, deep learning methods, especially long short-term memory (LSTM) networks, are widely used in human activity recognition [26–29]. However, they require big training data and high-dimensional feature vectors to perform well in classification tasks.

In previous study [10], LSTM showed dramatically the worst performance with low dimensional feature vectors among the other machine learning classifiers, which are not deep learning-based methods. In order to boost the performance of LSTMs in low dimensional feature space, in this paper, a novel scale and rotation invariant human activity recognition system, which employs LSTM network with low-dimensional 3D posture data, is presented. Since angles are used, proposed system is already scale invariant. In order to provide rotation invariance, body relative direction in egocentric coordinates is calculated. Different from the previous study [10], 3D joint angles are not employed as the feature vector. Instead, the angle of each limb vector with X, Y, and Z axes of the proposed egocentric coordinate system is employed as the feature vector. Additionally, several compression methods such as averaging filter, Haar wavelet transform (HWT), and discrete cosine transform (DCT) are employed to reduce dimension in feature vectors. This is an essential operation to attain real-time processing speed. Finally, RNN-LSTM network is employed to recognize five classes of human activities, namely, walking, standing, clapping, punching, and lifting. Experimental and benchmarking results show that proposed method dramatically (around 30%) increases the accuracy of LSTM classification in low

dimensional feature space compared to the previous method. The rest of the paper is organized as follows: Section 2 describes the methodology, Section 3 presents the experiment, experiment results, and evaluation; and conclusions are presented in Section 4.

2. Materials and Methods

Real-time human activity recognition systems require two main performance metrics, i.e., high accuracy and high processing speed. Since human activity recognition is a time series classification problem, LSTM's are well known for its excellent performance on time series classification. However, LSTM's have two disadvantages, i.e., requires high dimensional feature vectors and high number of instances in the training set. Even though it is possible to employ a big data set and high dimensional feature vectors in LSTM networks, the training and the processing speed (frame per second) of classification in real time may become dramatically low. In order to solve this problem, dimension reduction is an inevitable preprocessing stage to speed up the LSTM network. However, dimension reduction leads to loss of information, which causes low accuracy in the classification.

In previous study [10], we have experimented that LSTM's accuracy is dramatically lower than the other classifiers when the dimension reduction is applied to the feature vectors. Besides, rotation invariance has been achieved up to 90 degrees by training the users in different posture angles by providing 45 degrees of freedom in training session. Additionally, it requires too much time and effort to train the users, which is not automatic by the system and the posture angles are sometimes different on different users that creates lower performance in the classification. On the other hand, in previous study [10], 3D angles among the joints were employed as the feature vector, which are calculated with respect to global coordinate system's axes. Therefore, scale and rotation invariance could not be achieved truly, which caused low performance in LSTM classification with low dimensional feature vectors. The construction of rotation invariance in previous study [10] is illustrated in Figure 1.

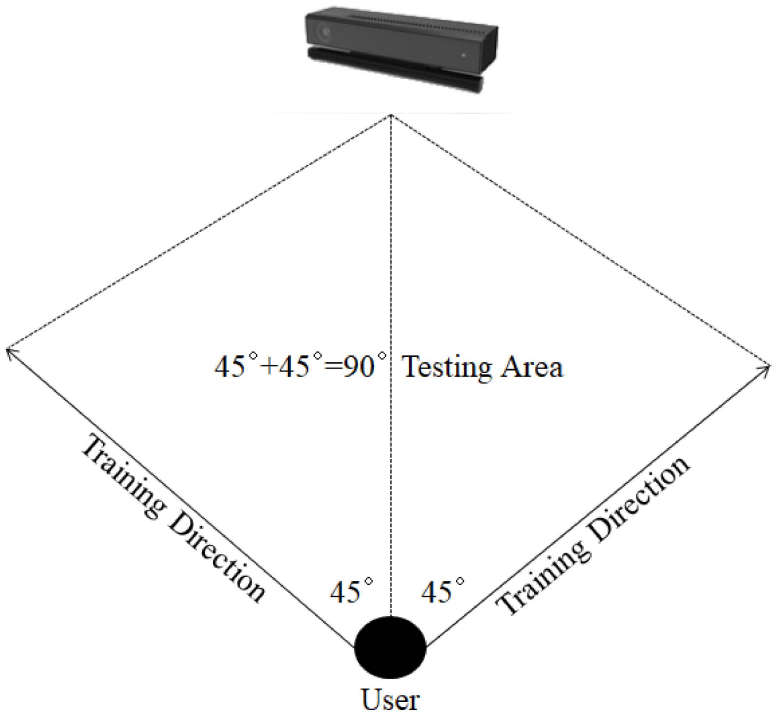

Figure 1. Construction of rotation invariance in previous study [10].

In order to solve these problems, in this study, which firstly presented briefly in [30], a scale and rotation invariant human activity recognition system based on body relative direction in egocentric coordinates is proposed. In the proposed system, Kinect depth sensor is employed to obtain skeleton

joints. Instead of using joint angles, the angle of each limb with X, Y, and Z axes of the proposed local coordinate system is employed as feature vector. Since angles are used, proposed system is already scale invariant. In order to provide rotation invariance, body relative direction in egocentric coordinates is calculated. The 3D vector between right hip and left hip is used to get the horizontal axis and its cross product with the vertical axis of global coordinate system is assumed to be the depth axis of the proposed local coordinate system.

As the system parameters, 8 number of limbs and their corresponding 3D angles with X, Y, and Z axes of the proposed coordinate system are employed as the feature vector. Since human activity recognition requires a period of time to recognize the action, n number of frames in a queue structure is employed as the period of action, which finally yields $8 \times 3 \times n$ features for each frame of the Kinect video. Even if the n = 10, which is assumed to be very small period, it creates 240 number of features at each frame of the video. Queue accumulation and formation of feature vectors are illustrated in Table 1.

Table 1. Queue accumulation and formation of feature vectors.

Frame No	Angle (π) Between Limb Vectors and Egocentric Axes			Class Label	Classification Status
#	X	Y	Z	–	–
1	0.93	0.13	0.54	–	Queue is not full
2	0.28	0.87	0.63	–	Queue is not full
3	0.87	0.06	0.71	–	Queue is not full
4	0.41	0.93	0.95	–	Queue is not full
5	0.17	0.22	0.25	–	Queue is not full
6	0.95	0.48	0.23	–	Queue is not full
7	0.71	0.55	0.49	–	Queue is not full
8	0.83	0.30	0.39	–	Queue is not full
9	0.15	0.05	0.43	–	Queue is not full
10	0.07	0.56	0.98	Standing	Classification between frames 1–10
11	0.42	0.01	0.29	Standing	Classification between frames 2–11
12	0.97	0.26	0.91	Lifting	Classification between frames 3–12
13	0.60	0.82	0.08	Lifting	Classification between frames 4–13
14	0.04	0.42	0.48	Punching	Classification between frames 5–14
15	0.42	0.51	0.36	Punching	Classification between frames 6–15
16	0.06	0.99	0.31	Clapping	Classification between frames 7–16
17	0.47	0.96	0.26	Clapping	Classification between frames 8–17
18	0.45	0.40	0.84	Walking	Classification between frames 9–18
19	0.06	0.68	0.78	Walking	Classification between frames 10–19

As seen in Table 1, every single limb vector forms 3 number of angles with the egocentric axes. Since 8 number of limb vectors are employed in the system, each frame creates $8 \times 3 = 24$ number of angles. Later, these 24 number of angles are accumulated to the employed queue data structure within the system. In Table 1, queue size is 10 and First-In-First-Out (FIFO) structure of the queue allows us easy allocation of data frame by frame. In other words, queue keeps the last n = 10 items in memory and this allows us instantaneous classification in each frame. Depending on the queue size, which is assumed to be n, system will store $8 \times 3 \times n$ number of angles within each frame after the queue is fully filled. Considering the case of n = 10, system will create $8 \times 3 \times 10 = 240$ number of features (angles) in each frame. This is a big number for real-time LSTM classification. Therefore, several dimension reduction methods such as averaging filter, Haar wavelet transform

(HWT), and discrete cosine transform (DCT) are applied to reduce the dimension size and eliminate the high-frequency noise. Flow chart of the proposed system is illustrated in Figure 2.

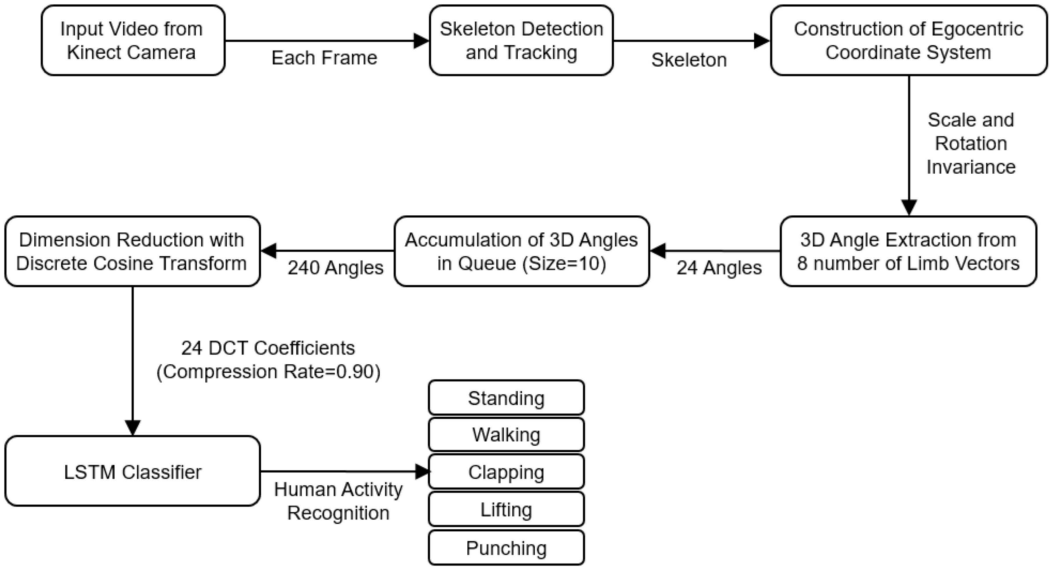

Figure 2. Flow chart of the proposed human activity recognition system.

In case of averaging filter, kernel size (queue size) determines the percentage of compression. If the total number of features in feature vector is 240 and kernel size (queue size) is 10, then, it reduces the number of features to 24, which is 90% compression. In case of HWT and DCT, low-frequency HWT and DCT coefficients are selected and high-frequency coefficients are discarded. In frequency domain, low frequency HWT and DCT coefficients are indexed towards the left while high frequency coefficients are indexed towards the right in one dimension. Therefore, a compression rate is employed as the system parameter to select the low-frequency coefficients. If the compression rate is 0.90, it means that the first 10% of the HWT and DCT coefficients are selected to be used as feature vector in the classifier. If the number of features is set to 240, the corresponding number of HWT and DCT coefficients is also 240, which is reduced to 24 when the compression rate is set to 0.90 in system settings. Finally, RNN-LSTM network is employed to recognize five classes of human activities, namely, walking, standing, clapping, punching, and lifting.

2.1. Egocentric Coordinate System Relative to Human Body Direction

In previous study [10], global coordinate system was used in which rotation invariance has been achieved up to 90 degrees by training the users in different posture angles by providing 45 degrees of freedom in training session. Additionally, it requires too much time and effort to train the users which is not automatic by the system and the posture angles are sometimes different on different users which creates lower performance in the classification.

In this study, an egocentric coordinate system, which is relative to human body direction, is presented. Similar with the previous study [10], Kinect depth sensor is employed to obtain skeleton joints. Since angles are used, proposed system is already scale invariant. In order to provide rotation invariance, body relative direction in egocentric coordinates is calculated. The 3D vector between right hip and left hip is used to get the horizontal axis, and its cross product with the vertical axis of global coordinate system is assumed to be the depth axis of the proposed local coordinate system. Different from the previous study [10], 3D joint angles are not being used, instead, 8 number of limbs and their corresponding 3D angles with X, Y, and Z axes of the proposed coordinate system are employed as the feature vector. In order to construct egocentric coordinate system, X axis is assumed as the normalized vector between the right hip and left hip. Additionally, Y axis is assumed as the vertical

unit vector of general coordinate system. Since horizontal and vertical axes are known, the depth axis Z is constructed by taking the cross product of these two vectors.

Proposed egocentric coordinate system, employed limb vectors, and constructed 3D vectors are illustrated in Figure 3, Figure 4, and Figure 5 as follows:

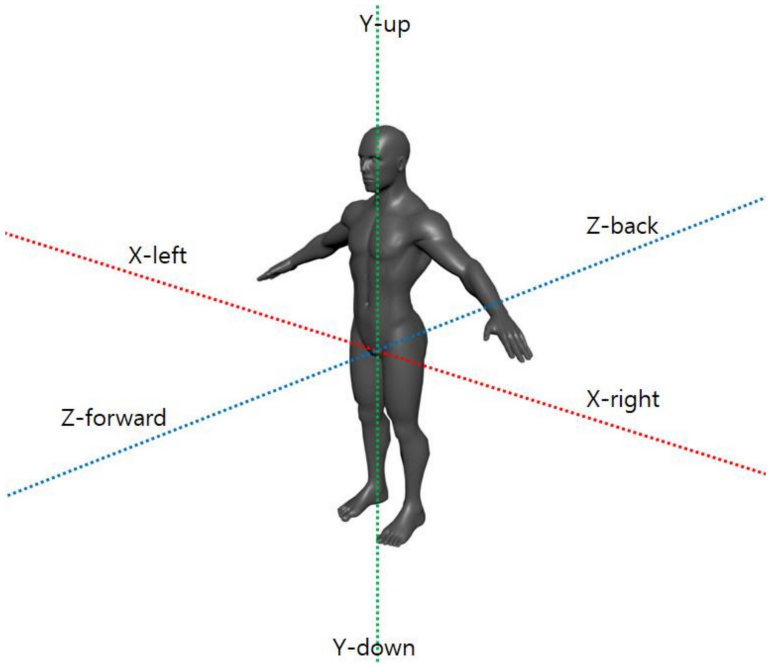

Figure 3. Egocentric coordinate system relative to human body direction.

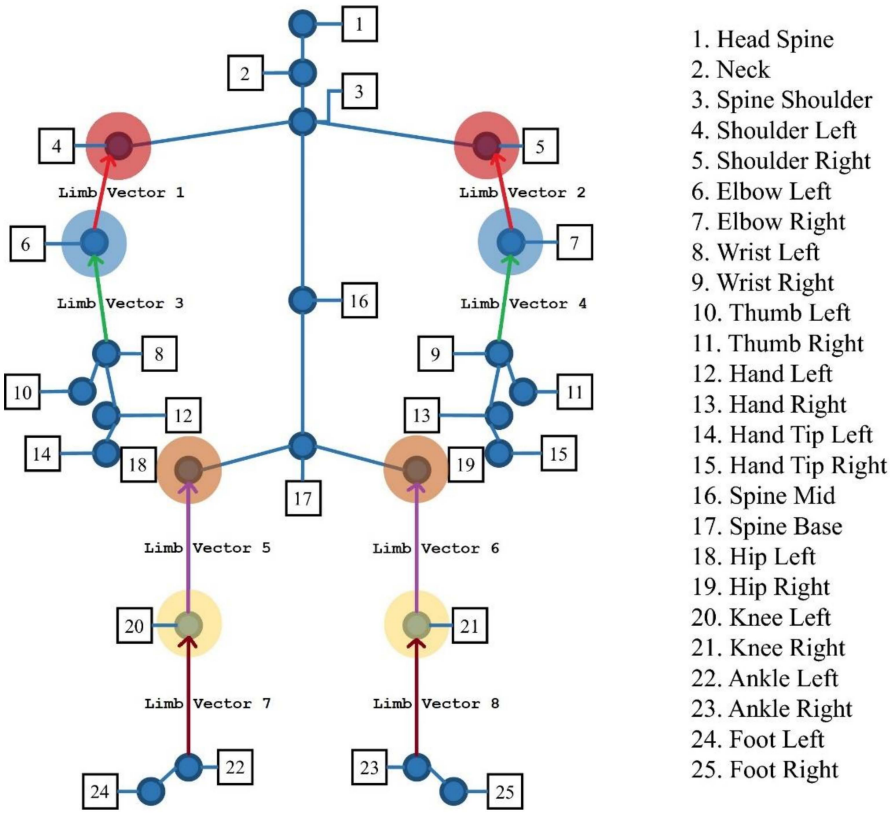

Figure 4. Extracted 8 number of limb vectors [10].

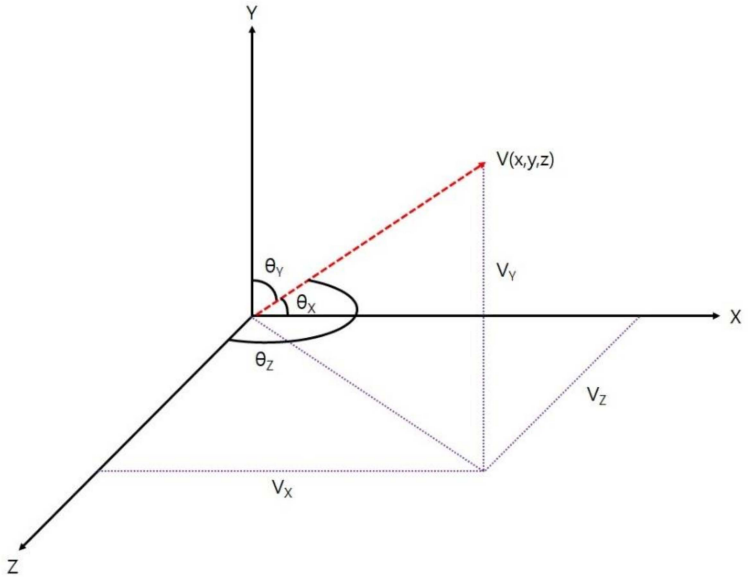

Figure 5. Limb vector and its 3D angles with X, Y, and Z axes in egocentric coordinate system.

Mathematically speaking, let $\hat{u}$ is an unit vector denoted as $\vec{u} = u\hat{u}$ and $\vec{u} = u_x\hat{i} + u_y\hat{j} + u_z\hat{k}$, where $\hat{i} = (1, 0, 0)$ is the unit vector along the X axis, $\hat{j} = (0, 1, 0)$ is the unit vector along the Y axis, and $\hat{k} = (0, 0, 1)$ is the unit vector along Z axis. The third vector, which is orthogonal to both $\vec{u}$ and $\vec{v}$, is found by the cross product of two vectors: $\left|\vec{u} \times \vec{v}\right| = uv \sin\theta$, where θ is the angle between $\vec{u}$ and $\vec{v}$. The calculation comes from the basic dot product formula shown below. The cosine angle between two vectors is actually the dot product of two vectors, which are normalized by dividing their components with the magnitude of each vector as follows:

$$\cos\theta = \frac{\left(\vec{u} \cdot \vec{v}\right)}{\left(\|\vec{u}\|\|\vec{v}\|\right)} \tag{1}$$

Since all the axes (X, Y, and Z) and limb vectors are normalized, it is easy to find the angle between each limb vector and X axis, Y axis, and Z axis separately. Finally, arccosine of the dot product gives the 3D angle between two vectors in a range of $[0, \pi]$.

2.2. Dimension Reduction with DCT

Discrete cosine transform (DCT) is a spatial to frequency domain transformation method, which represents a sequence of data in the form of a sum of cosine functions that oscillate at different frequencies. The DCT is an orthonormal transform in which $y = Cx$ and $x = C^{-1}y$ are defined in [31] as follows:

$$y(k) = \sqrt{\tfrac{2}{N}}\alpha(k) \sum_{n=0}^{N-1} x(n) \cos\tfrac{(2n+1)k\pi}{2N}; \quad k = 0,\ 1, \ldots N-1 \tag{2}$$

$$x(n) = \sqrt{\tfrac{2}{N}} \sum_{k=0}^{N-1} \alpha(k) y(k) \cos\tfrac{(2n+1)k\pi}{2N}; \quad n = 0,\ 1, \ldots N-1 \tag{3}$$

$$\alpha(0) = \frac{1}{\sqrt{2}};\ \alpha(k) = 1;\ k \neq 0. \tag{4}$$

Similar to Haar wavelet transform (HWT) and discrete Fourier transform (DFT), DCT concentrates the signal energy on a small number of low-frequency DCT coefficients. In the frequency domain,

compression (dimension reduction) is achieved by keeping the low-frequencyomponents (the most useful information) and discarding the high-frequency components that normally represent noise. Therefore, performing DCT can reduce the data size and noise level.

Table 2 demonstrates an example for the construction of DCT coefficients in frequency domain and indicates the lossy inverse transform as follows.

Table 2. Lossy discrete cosine transform (DCT) operations for sample inputs.

Input Vector	DCT Coefficients	Inverse DCT Transform
72	227.68398	71.99998
167	39.59834	167.00004
15	37.25152	14.99992
138	5.45163	138.00001
19	−31.71883	19.00003
43	−0.63242	42.99998
51	30.07613	51.00003
120	−89.54173	120.00009
10	−41.78117	9.99995
85	−111.25713	85.00000

In the above table, an input vector composed of 10 number of angle values ranging from 0 to π are passed through DCT (discrete cosine transform) and resulting DCT coefficients in frequency domain are illustrated. Additionally, in order to show how the inverse DCT is a lossy method, the resulting values after inverse DCT are also illustrated in the table.

According to the table, values after inverse DCT operation are not exactly same with the input values, which indicates that DCT is a lossy compression method. On the other hand, frequency levels are sorted in ascending order from low frequency to high frequency in the frequency domain. In this regard, dimension reduction is achieved by keeping the low-frequency information and discarding the high-frequency information in the frequency domain. In other words, instead of employing the input vector in spatial domain, low-frequency DCT coefficients in frequency domain are used as the feature vector. By this method, the most useful information is kept and unnecessary information is discarded, which reduces not only the feature dimension but also the noise level at the same time.

2.3. Deep Learning with RNN-LSTM Network

Different from the conventional neural network, LSTM (long short-term memory) network is designed to learn long-term interactions and recall information for long period of time by avoiding the long-term dependency problem. It was first proposed in [32] with a unique four layered communication structure, which consists of blocks of memory called cells where two number of states are transmitted to the next cell as the cell state and the hidden state. The cell state is the basic element of data stream, which provides forward transmission with little change in the data due to some linear transformations. Additionally, data can be manipulated (adding or removal of data) from the cell state by means of the sigmoid gates, which are actually designed as the series of matrix operations with varying individual weights. On the other hand, LSTMs are well known for their excellent performance on time series classification. LSTMs are capable of learning long-term dependencies and also prevent back-propagated errors from vanishing or exploding, thus avoiding the vanishing gradient problem.

Recurrent Neural Networks (RNNs) have the capability to capture temporal information from both sequential and spatial sequential data. Therefore, RNN-based LSTMs can simulate long windows of activity by replacing RNNs with LSTM storage locations [33,34]. The downside of RNNs is the problem of gradient explosion and decay, which interferes with the network's ability to model the

wide temporal relationships between input from a long contextual window and human activity [35]. The structure of the LSTM neural network is shown in Figure 6 as follows.

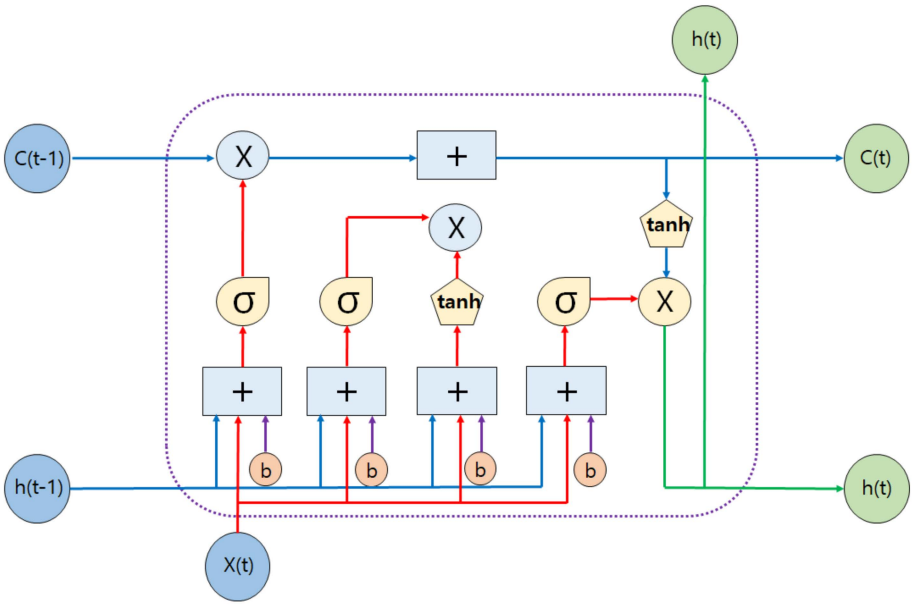

Figure 6. Structure of the long short-term memory (LSTM) neural network.

In the above figure, X_t is current input, C_{t-1} is memory from the LSTM unit, h_{t-1} is output of last LSTM unit, C_t is new updated memory, h_t is current output, σ is sigmoid layer, *tanh* is *Tanh* layer, b is bias, $\times$ is scaling of information, and $+$ is adding information. The equations of a typical LSTM network are given below:

$$f_t = \sigma\left(W_f[h_{t-1}, X_t] + b_f\right) \tag{5}$$

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \tag{6}$$

$$N_t = tanh(W_n[h_{t-1}, X_t] + b_n) \tag{7}$$

$$C_t = C_{t-1}f_t + N_t i_t. \tag{8}$$

Sigmoid function σ is employed for determining which information is not necessary and should be eliminated from the network. Therefore, it gets two arguments as the old output h_{t-1} at time $t-1$ and new input X_t at time t. Forget gate f_t, which is a vector of values between 0 and 1, is employed to decide whether the old output should be modified or partially eliminated for each cell state C_{t-1} with weight matrices W_f and bias b_f. For each new input, sigmoid layer and *tanh* function are employed to determine the importance of information by giving 1 for update operation and 0 for ignorance. Additionally, values to be updated are quantized with weights between -1 and 1 depending on their level of importance. Finally, new cell state C_t is updated from the old cell state C_{t-1} using the prior information obtained from the network.

LSTM's have two disadvantages, i.e., requires high dimensional feature vectors and high number of instances in the training set. Even though it is possible to employ a big data set and high dimensional feature vectors in LSTM networks, the training and the processing speed (frame per second) of classification in real time may become dramatically low.

3. Experimental Results

Different from the previous study's dataset [10], the dataset for this research was recreated by Kyungsung University, Department of Electronics Engineering since previous study's dataset does not include global coordinate system's axes information, which is a must to create our proposed model.

Although there exists plenty of publicly available datasets [36–39], none of them includes the global axes' coordinates, which is an obligatory information to create our proposed egocentric coordinate system. For the coding of the proposed system, previous study's C# code [10] was updated with the proposed method's implementation. For this purpose, Microsoft Visual Studio 2019 was chosen as the C# coding editor. Additionally, several external libraries such as Microsoft Kinect SDK 2.0, Vitruvius, and Accord.NET were used. On the other hand, Python 3.8.3 with TensorFlow was also used for the LSTM network performance evaluation.

The dataset created for this study contains information regarding 10 number of users who differ in height, weight, and clothing. Each activity for each person was logged twice to create a training set and once to create a test set. This is because the test sequences contain different angle variations of the same activity that are used to judge the accuracy of the proposed system under real-life conditions. RNN-LSTM model was employed on the dataset generated from the proposed coordinate system, which was separated into 70:30 for training and testing for the k-fold cross-validation. For the k-fold cross-validation, k was selected as 10. Since each person runs the activity at a different pace, the number of instances for each activity is different. Number of instances in the training and testing datasets employed in our experiments are listed in Table 3.

Table 3. Number of instances in the training and testing datasets.

Activity	Number of Instances (Training Dataset)	Number of Instances (Testing Dataset)
Standing	3355	1409
Walking	3388	1570
Clapping	3619	1397
Lifting	3645	1339
Punching	3476	1543
Total	17,483	7258

A sample snapshot from experimental environment is demonstrated in Figure 7 as follows.

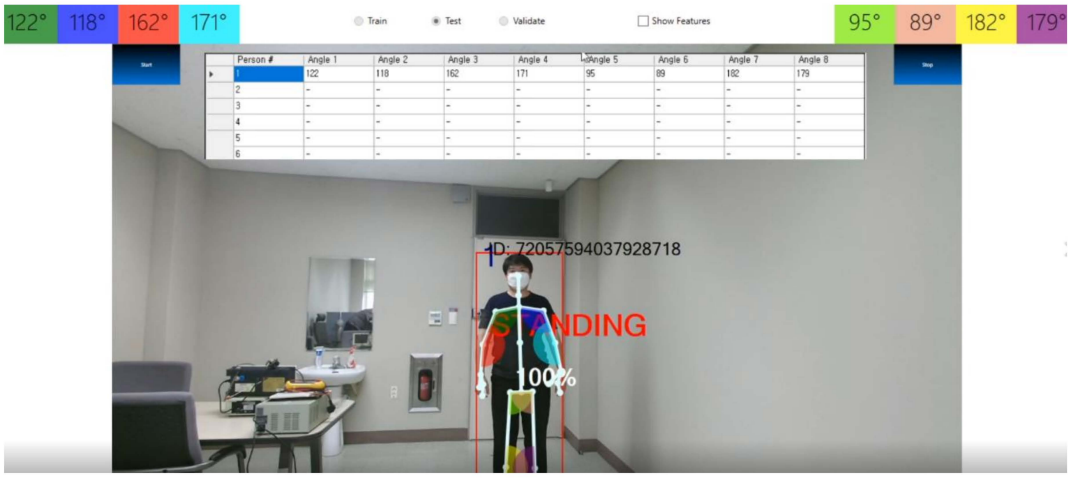

Figure 7. A snapshot from C#-based experimental environment.

In previous study [10], we have experimented that LSTM's accuracy is dramatically lower than the other classifiers when the dimension reduction is applied to the feature vectors. Although high-dimensional feature vectors can be used in LSTM networks, the training duration may become high and classification speed (frames per second) can be significantly low in real time. In order to

solve this problem, dimension reduction is an inevitable preprocessing stage to speed up the LSTM network. However, dimension reduction leads to loss of information which causes low accuracy in the classification.

In this study, an egocentric coordinate system is presented to boost the performance of LSTM in low-dimensional feature space. For this purpose, in the experimental setup, RNN-LSTM network is employed to recognize five classes of human activities, namely, walking, standing, clapping, punching, and lifting. Several compression methods such as averaging filter, Haar wavelet transform (HWT), and discrete cosine transform (DCT) are employed to reduce dimension in feature vectors. This is an essential operation to attain real-time processing speed. Besides, the effect of queue size on the performance of LSTM classification is observed with varying values. Table 4 summarizes the experiment's selected parameters as follows.

Table 4. Selected parameters for system settings.

Category	Parameter	Value
Setup	Time step	10
	Window size	100
	Batch size	64
	Epochs	75
Design	Hidden layers	32
	Neurons	30
Training	Activation function	Soft-max
	Bias weight	1.0
Learning	Optimizer	Adam
	Learning rate	0.0025
	Loss rate	0.0015

Additionally, in case of employing DCT compression rate of 90% (24 number of features), Figure 8 shows the graph of accuracy for both training and testing cost. The confusion matrix obtained after the cross-validation is presented in Figure 9. Besides, employed performance metrics are listed in Table 5.

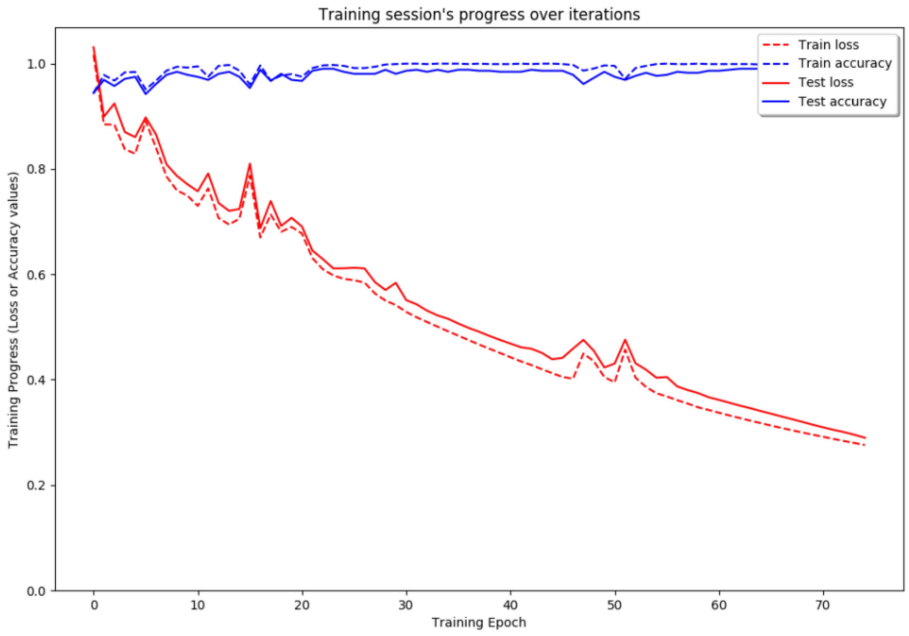

Figure 8. Training and testing progress over 75 epochs.

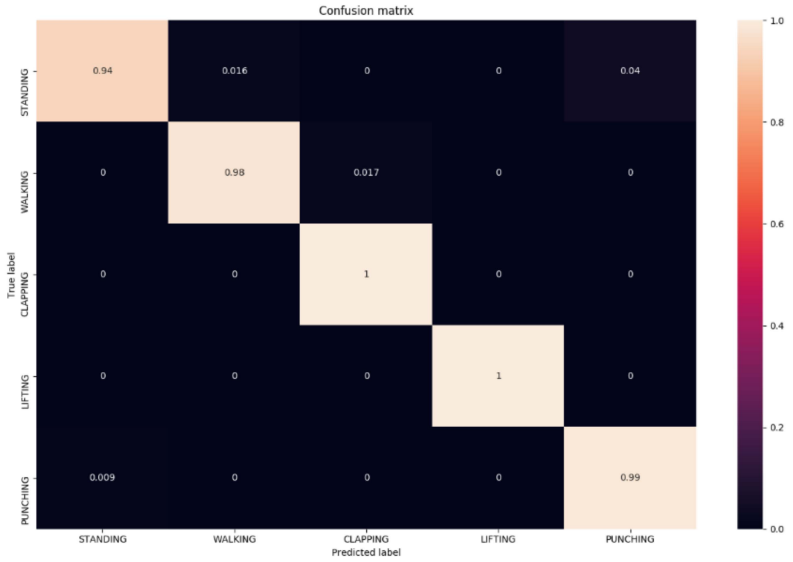

Figure 9. Confusion matrix.

Table 5. Performance metrics employed in the experiments.

Abbreviation	Description
TPR	Sensitivity or Recall or True Positive Rate
TNR	Specificity or True Negative Rate
PPV	Precision or Positive Predictive Value
NPV	Negative Predictive Value
FPR	False Positive Rate or Fall-out
FNR	False Negative Rate or Miss Rate
FDR	False Discovery Rate
FOR	False Omission Rate
ACC	Accuracy
F1	F-measure or F1 Score

In addition, experimental and benchmarking results are shown in Table 6, Table 7, and Table 8. Finally, the relationship between the LSTM's accuracy and DCT vector size (n) is listed in Table 9 and illustrated with a graph in Figure 10, and evaluation of the experimental and benchmarking results is done at the end.

Table 6. Cross-validation results of LSTM classifier based on the employed performance metrics.

Metrics/Classes	Standing	Walking	Clapping	Lifting	Punching	Average
TPR	62.667	85.217	100.00	100.000	91.667	87.910
TNR	99.775	99.469	99.157	100.000	99.899	99.660
PPV	91.262	85.965	5.556	100.000	96.117	75.780
NPV	98.615	99.436	100.00	100.000	99.773	99.565
FPR	0.225	0.531	0.843	0.000	0.101	0.340
FNR	37.333	14.783	0.000	0.000	8.333	12.090
FDR	8.738	14.035	94.444	0.000	3.883	24.220
FOR	1.385	0.564	0.000	0.000	0.227	0.435
ACC	98.432	98.946	99.157	100.000	99.68	99.243
F1	74.308	85.59	10.526	100.000	93.839	72.853

Table 7. Benchmarking of cross-validation results of LSTM classifier with different feature vectors.

Feature Vector	Accuracy (%)	F-1 (%)	Precision (%)	Recall (%)
Dimension reduction with averaging (Previous Coordinate System)	65.7	73.9	78.1	67.9
Dimension reduction with averaging (Proposed Coordinate System)	85.1	65.4	68.3	71.5
Dimension reduction with HWT (Previous Coordinate System)	75.7	79.3	81.6	74.3
Dimension reduction with HWT (Proposed Coordinate System)	93.5	68.7	71.3	80.2
Dimension reduction with DCT (Previous Coordinate System)	83.1	85.4	87.6	82.7
Dimension reduction with DCT (Proposed Coordinate System)	99.2	72.9	75.8	87.9

Table 8. Benchmarking of experimental results of LSTM classifier with different feature vectors.

Feature vector	Accuracy (%)	F-1 (%)	Precision (%)	Recall (%)
Dimension reduction with averaging (Previous Coordinate System)	53.8	48.1	53.7	45.3
Dimension reduction with averaging (Proposed Coordinate System)	68.1	62.9	65.2	69.0
Dimension reduction with HWT (Previous Coordinate System)	57.1	56.1	59.4	55.1
Dimension reduction with HWT (Proposed Coordinate System)	72.5	63.6	66.1	71.6
Dimension reduction with DCT (Previous Coordinate System)	61.6	63.4	68.2	62.6
Dimension reduction with DCT (Proposed Coordinate System)	83.9	66.7	71.4	74.5

Table 9. Benchmarking of accuracies of previous and proposed method in varying DCT vector sizes.

DCT Vector Size	Accuracy (%) (Previous Method)	Accuracy (%) (Proposed Method)
1	5.3	17.9
2	17.1	24.5
4	25.4	31.8
8	32.7	45.2
12	41.6	68.3
16	49.7	75.5
20	55.1	79.4
24	61.6	83.9
30	64.3	85.4
40	65.7	87.2
48	66.3	88.1
60	67.8	89.4
80	68.5	90.1
120	69.3	90.5
240	70.2	91.7

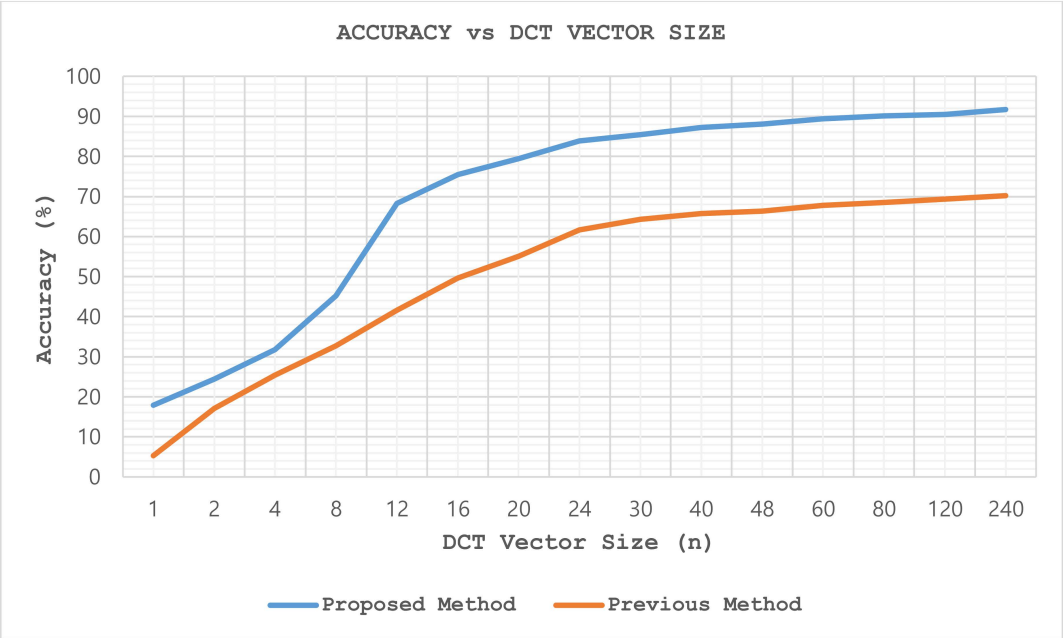

Figure 10. Graph of accuracies of previous and proposed method in varying discrete cosine transform (DCT) vector sizes.

According to experimental results, DCT yields the highest accuracy among other dimension reduction methods. As expected, HWT yields lower accuracy than DCT and gives higher accuracy than averaging filter. Both cross-validation with training data and performance measurement with testing data indicate that accuracy is increasing by the increment of feature vectors' size. Additionally, increment slope smoothly decreases by the increment of feature vectors' size. However, an optimum threshold in terms of compression rate is required in order to judge the system whether it is applicable in real time. Therefore, the compression rate of 90% (240 number of features are reduced to 24) is chosen as the optimum value of compression, which yields 83.9% accuracy while it is being used with DCT. Here, it is assumed that accuracy greater than 80% is applicable and reasonable in real-time applications. On the other hand, the compression rate can be lowered to get higher accuracies depending on the performance of hardware in which the HAR system is running. If the performance of hardware is high enough to process the LSTM classification in real time without dimension reduction, proposed method yields 91.7% accuracy without dimension reduction, whereas previous method can achieve 70.2% accuracy only. This also proves that proposed method is dramatically (around 30%) better than previous method in terms of keeping higher information in feature vectors, which results in yielding higher accuracies in all the cases.

4. Conclusions

Human activity recognition (HAR) is a common time series classification task, which requires high accuracy with high processing speed in real-time applications. LSTM networks are widely used in time series classification problems, whereas they require big training data and high-dimensional feature vectors to get optimum performance, which dramatically increase the training duration and reduces the processing speed. Dimension reduction methods are generally employed to process LSTMs in low dimensional feature space, which usually yields low performance. In previous study [10], LSTM showed dramatically the worst performance with low dimensional feature vectors among the other machine learning classifiers, which are not deep learning-based methods. The reason for that was the discrimination power of the feature vectors constructed using the 3D joint angles in global coordinate system was very weak to get high accuracy in LSTM network. In order to boost the performance of LSTMs in low dimensional feature space, in this paper, a novel egocentric

coordinate system is presented. Based on the body relative direction of the users in the camera vision, proposed method provides a scale and rotation invariant human activity recognition system, which employs LSTM network with low-dimensional 3D posture data. In the proposed framework, Kinect depth sensor is employed to obtain skeleton joints. Since angles are used, proposed system is already scale invariant. In order to provide rotation invariance, body relative direction in egocentric coordinates is calculated. The 3D vector between right hip and left hip is used to get the horizontal axis and its cross product with the vertical axis of global coordinate system assumed to be the depth axis of the proposed local coordinate system. Instead of using 3D joint angles, 8 number of limbs and their corresponding 3D angles with X, Y, and Z axes of the proposed coordinate system are employed as the feature vector. In terms of dimension reduction, averaging filter, HWT (Haar wavelet transform) and DCT (discrete cosine transform) are employed with varying kernel sizes. Sliding kernel's functionality is achieved using a specific queue data structure. Finally, extracted features are trained and tested with LSTM (long short-term memory) network which is an artificial recurrent neural network (RNN) architecture. Experimental results indicate that DCT compression has the minimum loss of information among other dimension reduction methods and proposed framework dramatically increases the discrimination power of feature vectors. Using the proposed egocentric coordinate system, LSTM achieves outstanding results with 83.9% accuracy with an optimum DCT compression rate of 90%. Additionally, a benchmarking study is performed with the previous study's method [10] in which the highest accuracy is obtained with 61.6% where rotation invariance is satisfied while rotating by the 45 degrees of freedom in training session. Benchmarking results show that proposed method overwhelms the previous method dramatically (approximately 30% in accuracy) and yields excellent results. As the future work, other attempts to obtain different source of coordinate input will be tested instead of using the Kinect camera, e.g., such as from CCTV real time video without a depth sensor.

Funding: This research received no external funding.

Acknowledgments: This research was supported by the Kyungsung University Research Grants in 2019 and Brain Busan 21+ 2020.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Lin, W.; Sun, M.T.; Poovandran, R.; Zhang, Z. Human activity recognition for video surveillance. In Proceedings of the 2008 IEEE International Symposium on Circuits and Systems, Seattle, WA, USA, 18 May–21 August 2008; pp. 2737–2740. [CrossRef]
2. Sebestyen, G.; Stoica, I.; Hangan, A. Human activity recognition and monitoring for elderly people. In Proceedings of the 2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 8–10 September 2016; pp. 341–347. [CrossRef]
3. Jalal, A.; Zeb, M.A. Security Enhancement for E-learning portal. *Int. J. Comput. Sci. Netw. Secur.* **2008**, *3*, 41–45.
4. Tentori, M.; Favela, J. Activity-aware computing for healthcare. *Pervasive Comput. IEEE* **2008**, *7*, 51–57. [CrossRef]
5. Jalal, A.; Zeb, M.A. Collaboration achievement along with performance maintenance in video streaming. In Proceedings of the International Conference on Computer and Information Technology, Dhaka, Bangladesh, 27–29 December 2007; pp. 369–374.
6. Jalal, A.; Kamal, S.; Kim, D. A depth video-based human detection and activity recognition using multi-features and embedded hidden Markov models for health care monitoring system. *Int. J. Interact. Multimed. Artif. Intell.* **2017**, *4*, 54–62. [CrossRef]
7. Subasi, A.; Radhwan, M.; Kurdi, R.; Khateeb, K. IoT based mobile healthcare system for human activity recognition. In Proceedings of the Learning and Technolgy Conference (L&T), Jeddah, Saudi Arabia, 25–26 February 2018; pp. 29–34.
8. Kamal, S.; Jalal, A.; Kim, D. Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM. *J. Electr. Eng. Technol.* **2016**, *6*, 1857–1862. [CrossRef]

9. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recogn. Lett.* **2019**, *119*, 3–11. [CrossRef]
10. Ince, O.F.; Ince, I.F.; Yildirim, M.E.; Park, J.S.; Song, J.K.; Yoon, B.W. Human activity recognition with analysis of angles between skeletal joints using a RGB-depth sensor. *ETRI J.* **2019**, *42*, 2–3. [CrossRef]
11. Koller, D.; Klinker, G.; Rose, E.; Breen, D.; Whitaker, R.; Tuceryan, M. Real-time vision-based camera tracking for augmented reality applications. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology, Lausanne, Switzerland, 15–17 September 1997; pp. 87–94.
12. Jalal, A.; Kamal, S. Real-time life logging via a depth silhouette-based human activity recognition system for smart home services. In Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, Seoul, Korea, 26–29 August 2014; pp. 74–80.
13. Al Machot, F.; Elkobaisi, M.R.; Kyamakya, K. Zero-Shot Human Activity Recognition Using Non-Visual Sensors. *Sensors* **2020**, *20*, 825. [CrossRef] [PubMed]
14. Ding, M.; Fan, G. Articulated and generalized Gaussian kernel correlation for human pose estimation. *IEEE Trans. Image Process.* **2016**, *25*, 776–789. [CrossRef] [PubMed]
15. Ye, M.; Yang, R. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 2345–2352.
16. Hbali, Y.; Hbali, S.; Ballihi, L.; Sadgal, M. Skeleton-based human activity recognition for elderly monitoring systems. *IET Comput. Vis.* **2018**, *12*, 16–26. [CrossRef]
17. Shotton, J.; FitzGibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time human pose recognition in parts from single depth images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 119–135.
18. Jalal, A.; Quaid, M.A.K.; Hasan, A.S. Wearable sensor-based human behavior understanding and recognition in daily life for smart environments. In Proceedings of the International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 17–19 December 2018; pp. 105–110.
19. Sikder, F.; Sarkar, D. Log-sum distance measures and its application to human-activity monitoring and recognition using data from motion sensors. *IEEE Sensors* **2017**, *14*, 4520–4533. [CrossRef]
20. Luo, X.; Tan, H.; Guan, Q.; Liu, T.; Zhuo, H.H.; Shen, B. Abnormal activity detection using pyroelectric infrared sensors. *Sensors* **2016**, *16*, 822. [CrossRef] [PubMed]
21. Chen, Y.; Shen, C. Performance analysis of smartphone-sensor behavior for human activity recognition. *IEEE Access* **2017**, *5*, 3095–3110. [CrossRef]
22. Nguyen, T.N.; Ly, N.Q. Abnormal activity detection based on dense spatial-temporal features and improved one-class learning. In Proceedings of the Eighth International Symposium on Information and Communication Technology-SoICT, Nha Trang City, Vietnam, 7–8 December 2017; pp. 370–377.
23. Singh, D.; Mohan, C.K. Graph formulation of video activities for abnormal activity recognition. *Pattern Recogn.* **2017**, *65*, 265–272. [CrossRef]
24. Mahmood, M.; Jalal, A.; Sidduqi, M.A. Robust spatio-temporal features for human interaction recognition via artificial neural network. In Proceedings of the 2018 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 17–19 December 2018; pp. 218–223.
25. Sharif, M.; Khan, M.A.; Zahid, F.; Shah, J.H.; Akram, T. Human action recognition: A framework of statistical weighted segmentation and rank correlation-based selection. *Pattern Anal. Appl.* **2020**, *23*, 281–294. [CrossRef]
26. Wang, K.; Wang, X.; Lin, L.; Wang, M.; Zuo, W. 3D human activity recognition with reconfigurable convolutional neural networks. In Proceedings of the ACM International Conference on Multimedia-MM'14, Orlando, FL, USA, 3–7 November 2014; pp. 97–106.
27. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
28. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
29. Tao, D.; Wen, Y.; Hong, R. Multicolumn bidirectional long short-term memory for mobile devices-based human activity recognition. *IEEE Internet Things J.* **2016**, *3*, 1124–1134. [CrossRef]

30. Wesonga, S.; Furkan, I.I.; Park, J.-S. Scale and Rotation Invariant Human Activity Recognition based on Body Relative Direction in Egocentric Coordinates. In Proceedings of the International Conference on Control, Automation and Systems, Seoul, Korea, 13–16 October 202; pp. 395–397.
31. Ahmed, N.; Natarajan, T.; Rao, K.R. Discrete cosine transform. *IEEE Trans. Comput.* **1974**, *100*, 90–93. [CrossRef]
32. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
33. Agarwal, P.; Alam, M. A Lightweight Deep Learning Model for Human Activity Recognition on Edge Devices. *Procedia Comput. Sci.* **2020**, *167*, 2364–2373. [CrossRef]
34. Sagha, H.; Digumarti, S.T.; Millán, J.D.R.; Chavarriaga, R.; Calatroni, A.; Roggen, D.; Troster, G. Benchmarking classification techniques using the opportunity human activity dataset. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Anchorage, AK, USA, 9–12 October 2011; pp. 36–40.
35. Zhao, Y. Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors. *Math. Probl. Eng.* **2018**, *2018*, 7316954. [CrossRef]
36. Morana, M.; Lo Re, G.; Gaglio, S. KARD-Kinect Activity Recognition Dataset. *Mendeley Data* **2017**, *1*. [CrossRef]
37. Cornell Activity Datasets: CAD-60 & CAD-120. Available online: https://www.re3data.org/repository/r3d100012216 (accessed on 16 November 2020).
38. Xia, L.; Chen, C.-C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3D joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 20–27.
39. Seidenari, L.; Varano, V.; Berretti, S.; del Bimbo, A.; Pala, P. Recognizing Actions from Depth Cameras as Weakly Aligned Multi-Part Bag-of-Poses. In Proceedings of the 3rd International Workshop on Human Activity Understanding from 3D data (HAU3D'13), in conjunction with CVPR 2013, Portland, OR, USA, 24 June 2013.