Received: 26 August 2021 | Revised: 2 October 2021 | Accepted: 2 October 2021
DOI: 10.37917/ijeee.17.2.20 Vol. 17| Issue 2| December 2021

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Human Activity Recognition Using The Human Skeleton Provided by Kinect

Heba A. Salim*, Musaab Alaziz, Turki Y. Abdalla
Department of Computer Engineering, University of Basrah, Iraq

Correspondence
* Heba A. Salim
Department of Computer Engineering,
University of Basrah, Iraq
Email: pgs2343@uobasrah.edu.iq

Abstract
In this paper, a new method is proposed for people tracking using the human skeleton provided by the Kinect sensor, Our method is based on skeleton data, which includes the coordinate value of each joint in the human body. For data classification, the Support Vector Machine (SVM) and Random Forest techniques are used. To achieve this goal, 14 classes of movements are defined, using the Kinect Sensor to extract data containing 46 features and then using them to train the classification models. The system was tested on 12 subjects, each of whom performed 14 movements in each experiment. Experiment results show that the best average accuracy is 90.2 % for the SVM model and 99 % for the Random forest model. From the experiments, we concluded that the best distance between the Kinect sensor and the human body is one meter.

KEYWORDS: Classification, Kinect skeleton, People tracking, Random Forest, Support Vector Machine (SVM).

I. INTRODUCTION

Human activity recognition is one of the most widely researched topics, with numerous applications including human-computer interface, smart video surveillance, sports, and health care [1], [2]. In this work, the main aim is to recognize human activity by using the skeleton data provided by the Kinect sensor [3]. The benefits of using Kinect devices include the ability to track the position of human body joints in low light conditions, being suitable for the indoor environment's location, being reliable, and having a low cost. By capturing skeleton data, the Kinect sensor can recognize human activities. The skeleton data can provide the coordinate value of each joint detected by the Kinect. Three sorts of data are available from the Kinect sensor [4]:

• RGB images: Provided by the RGB camera. The Kinect uses this data to learn about the objects and people in the room [5], and it works like a 2D camera.
• Depth data: are provided by the depth sensor which is the combination of the IR projector and the depth camera, that gives the distance between the human's body and the depth camera.
• Skeleton data: the traced skeleton provided by the Kinect consist of twenty joint positions, each joint have three coordinates (x, y, z) [5].

The skeleton data is used to classify human activity based on their movement [6]. 14 activities have been chosen; standing in the room, prayer position, sitting on the floor with stretched legs, sitting on the floor with crossed legs, lying on the floor, lying on the floor with one leg raised, sitting on the bed, standing on the bed, standing on the bed with stretched legs, lying the bed with one leg raised, lying on the bed, sitting on the chair, sitting on the chair with crossed legs (the right leg on the left leg), sitting on the chair with crossed legs (the left leg on the right leg). These activities will be used to build our dataset for the training and testing process. The data will be classified using two methods the Support Vector Machine (SVM) which is chosen because it performs best in a large feature space, can handle large amounts of data better, and is suitable for classifying non-linear data. The other used method is the Random Forest algorithm, which consists of a large number of decision trees The random forest algorithm generates a 'forest,' which is trained using bagging or bootstrap aggregation. Bagging is a meta-algorithm ensemble that improves the accuracy of machine learning algorithms, experiments showed that these methods give a very good performance. The proposed tracking scheme is the main contribution of this work that we provide.

The remainder of the paper is organized as follows. In Section II, the related works are listed; in Section III, the methodology with the classification methods and feature extraction are listed in this section. In Section IV, all experiments are described and the results are presented in Section V presented. Finally, in Section VI, concluding remarks and feature works are provided.

II. Related Work

Many techniques for human activity recognition have been proposed in recent years, with some of them aiming to extract features from depth data, also they used fewer features and body positions in their work, these papers are as follows:

E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante et al. [7] Have proposed an algorithm for activity recognition. This algorithm is based on the skeleton data extracted from an RGBD sensor and creates a feature vector to represent the whole activity. The proposed algorithm works with two publicly available datasets, the KARD and CAD-60. Then the researchers R. Vemulapalli, F. Arrate, and R. Chellappa, et al. [8] Introduced R3DG features, which are a set of 3D skeleton representations of body parts for human action recognition. The proposed models explicitly model the two objects' relative 3D geometry. Rigid-body transformations are used in various body sections that are not directly connected by a joint. T. H. An, T. Q. Phuc, N. T. Hai, and T. T. Mai, et al. [9] In this study, researchers reported on human fall recognition using bone data. Using the Kinect camera system, they conduct three distinct experiments with a database build from the human skeleton. To identify fall cases and daily activities, the SVM algorithm was used, and it has great accuracy. Furthermore, this technique demonstrated the greatest promise for using skeletal data to recognize human poses.S. Gaglio, G. Lo Re, and M. Morana, et al. [10] Propose a system for recognizing human activities from 3-D posture data. They specifically mentioned a scenario in which the entire environment is outfitted with several sensory nodes that can monitor raw metrics like temperature, humidity, and light level in an unobtrusive manner. In this case, the Kinect is in charge of acquiring high-level data regarding the user's actions. M. W. Rahman and M. L. Gavrilova, et al. [11] Introduced a method for identifying people based on sensor-based gait data. The project's goal is to identify a person by utilizing 3D skeletal joint gait data from the Kinect. Each individual's gait cycle is detected, and features are got to learn using a KNN classifier.

III. Methodology

We developed a system for human action recognition that classifies action names based on human movement using support vector machines and random forest algorithms. To obtain the classification result, supervised learning is used, which must include training data. Fig. 2 shows an overview of the purposed system.

A. Skeleton joints position

The Kinect sensor can track the human body with various joint points using skeleton tracking. Using the Kinect for Windows SDK, it can track up to six persons and up to 20 joints for each skeleton. Only two persons can be tracked in detail, which means the sensor can return information on all 20 tracked joint points, whereas reset persons only get the overall position. This is because tracking joint information for all six persons would require a large amount of processing [12].

This indicates that the sensor can track 20 skeletal joint points. That twenty joints are the head, shoulder center, spine, hip center, left hip, right hip, right shoulder, right elbow, right wrist, right hand, left shoulder, left elbow, left wrist, left hand, right knee, right ankle, right foot, left knee, left ankle, left foot. For the highest performance and precision, all joint positions are employed in this study. The figure below illustrates the skeleton joint position.

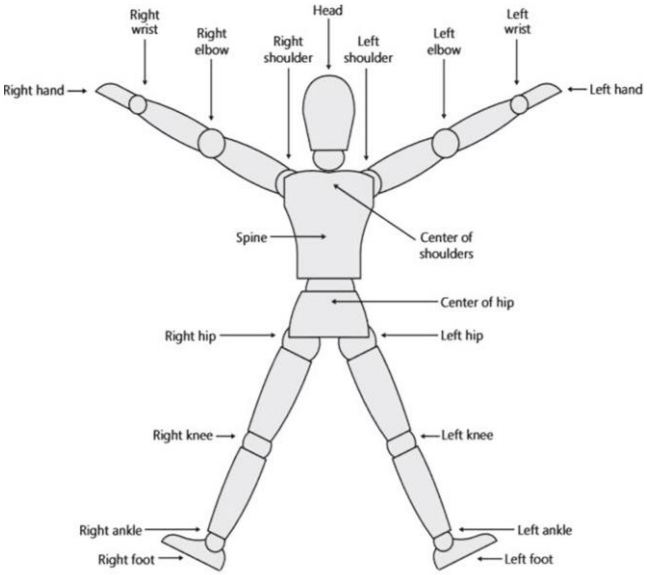

Fig. 1: Skeleton joints names.

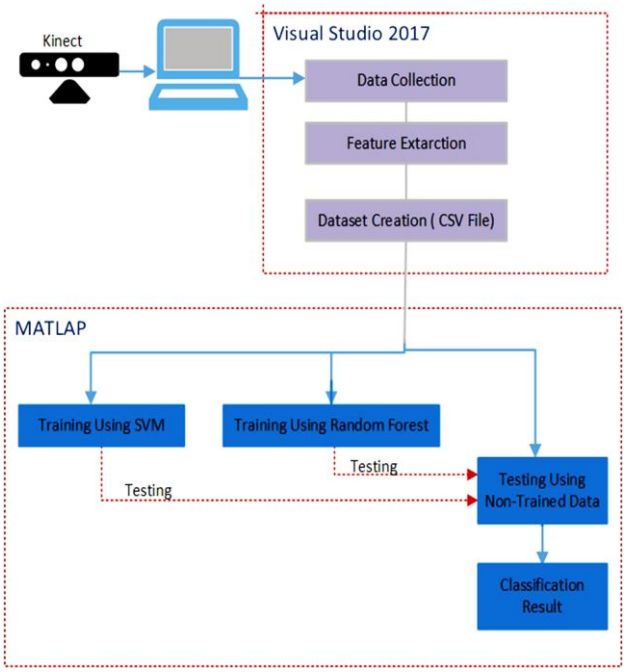

Fig. 2: Overview of system flow.

B. *Dataset Collection and Feature extraction*

First, extract the joint position from the skeleton monitored by the Kinect; each joint has three values of the (x, y, z) coordinates As shown in Fig. 3. [1]. Many properties can be obtained from these values, including joint angles, joint distances, distances between joints from the Kinect sensor, and distances between the Kinect and the human skeleton. There will be a total of 46 features. Then from these features, 14 body positions are driven as shown in TABLE I. Finally, the dataset was build based on these values. This dataset will be used for training and testing steps for the classification models.

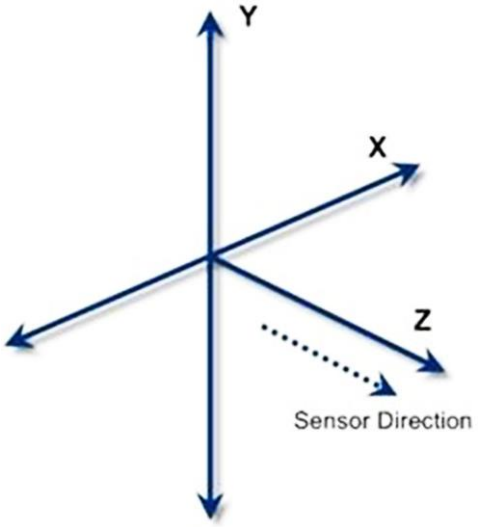

Fig. 3: Three coordinates for joint position [20].

C. *Classification methods*

In our work two classification methods have been used:

1) *Support vector machine*

Cortes and Vapnik developed the Support Vector Machine (SVM) method in 1995 to solve classification and regression problems. SVM's concept is to find the best line that divides two classes and determines the support vector. A hyperplane is a name given to this line. Support vector machines are typically used for linear classification. SVM, on the other hand, was designed to be used for non-linear classification by utilizing a kernel trick on feature space [1], [13]. The fundamental principle of SVM is that the function transfers data x to a vector space with a higher dimension (x) [14].

In this paper, the Gaussian Radial Basis Function kernel (RBF) was used. Regardless matter whether the sample is low-dimensional, high-dimensional, large-scale, or small-scale. The Gaussian kernel function is seen to be the best option because it has a larger convergence domain [3]. The RBF kernel's expression is,

$$K(\vec{x}_i, \vec{x}_j) = \exp\left(\frac{-\left|\left|\vec{x}_i - \sigma\vec{x}_j\right|\right|_2}{2\sigma^2}\right) \quad (1)$$

TABLE I
THE SELECTED CLASSES AND THEIR DESCRIPTION

Classe No.	Definition	
1.	Standing in the room	
2.	Sitting on the floor (the prayer position)	
3.	Sitting on the floor with stretched legs	
4.	Sitting on the floor with crossed legs	
5.	Lying the floor	
6.	Lying the floor with one leg raised	
7.	Sitting on the bed	
8.	Standing on the bed	
9.	Standing on the bed with stretched legs	
10.	Lying the bed with one leg raised	
11.	Lying the bed	
12.	Sitting on the chair	
13.	Sitting on the chair with crossed legs (The right leg on the left leg)	
14.	Sitting on the chair with crossed legs (The left leg on the right leg)	

The SVM model was trained it using our dataset, and the data is divided using cross-validation, with 50% of the data randomly selected for training and 50% for testing, and the procedure is repeated for 10 iterations, with the accuracy being evaluated at each iteration. After training the confusion matrix is determined as shown in Figure below.

As we can see in Fig. 4. The rows of the confusion matrix represent the true classes, while the columns of the matrix represent the predicted classes. The numbers in the blue cells represent correct predictions. The diagonal here is where the model performed accurately, and these cells will have high values in comparison to other cells in the same row. Because the true positives and true negatives are along the diagonal from top-left to bottom-right, we can assume the model is performing well if this diagonal is highlighted.

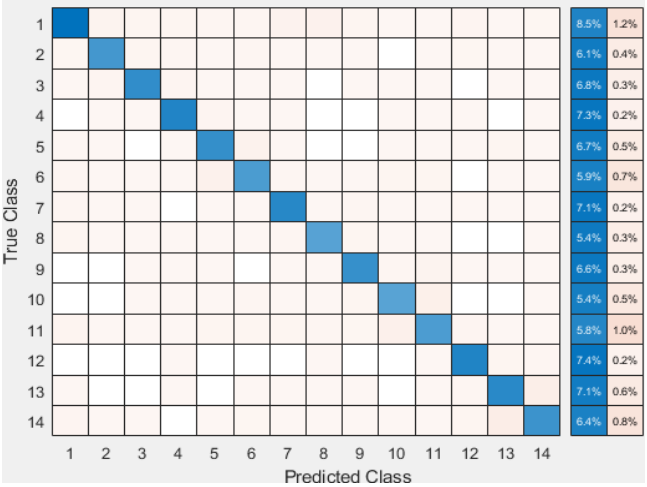

Fig. 4: The confusion matrix of the SVM model.

2) Random Forest

The random forests classifier, formed by a set of randomized decision trees, these trees are trained from the training sets. From the training sets, M segment features are randomly chosen and placed at a root node in each decision tree, that is connected to a group of terminating leaf nodes via the inner binary splitting joints [15]. At each interior joint, f variables from the F feature dimension are picked at random, and the decision threshold t is determined from the range $\{t|\text{mint } f(vi) \leq t \leq \text{maxi } f(vi)\}$. The definition of splitting function is:

$$f(vi) = \begin{cases} 1, & if \ \{i \in I | f(vi) > t\} \\ 0, & otherwise \end{cases} \quad (2)$$

The random forests model was trained with the important predictors determined from the "predictor importance function". To chose the best predictors we need to threshold them. So the threshold decided to be a value of "2" (see Fig. 5), then the number of important features will be 29 features which will be used in the training and testing steps. Using our dataset, and the data is divided using cross-validation, with 50% of the data randomly selected for training and 50% for testing, and the procedure is repeated for 10 iterations, with the accuracy being evaluated at each iteration.

After training the confusion matrix is determined as shown in Fig. 6. As we can see the numbers in the blue cells represent correct predictions. The diagonal is where the model has performed accurately. Because the true positives and true negatives are located along the diagonal from top-left to bottom-right, we can assume that the model is performing well if this diagonal is highlighted. As we can see the correct predicted values in this confusion matrix are higher than the values of the confusion matrix of the SVM model.

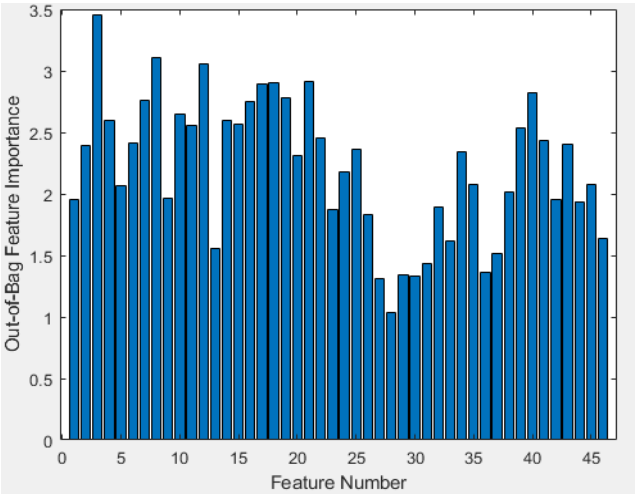

Fig. 5: Predictor importance of 46 featuer.

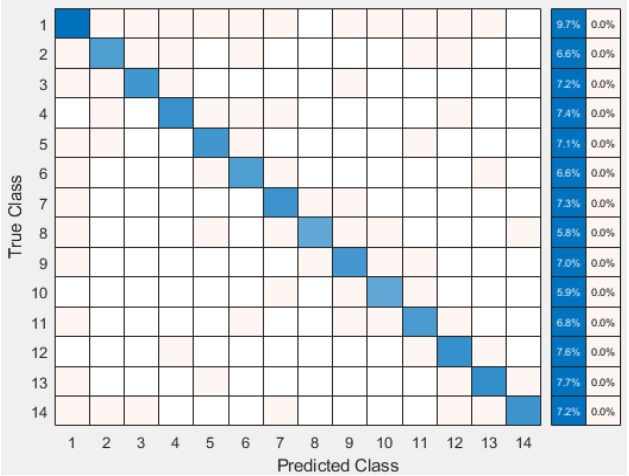

Fig. 6: The confusion matrix of the Random Forest model.

IV. Experiments

The experiments were repeated five times for each activity to determine which prediction was accurate and which was incorrect. The tests were carried out on a total of 12 people (9 men and 3 women). Because the Kinect sensor can not track the joints of the left hip, right hip, right knee,

right ankle, right foot, left knee, left ankle, and left foot, testing data on stand-up position have a high error over a 1-meter distance. The Kinect sensor and the human body are too close. The best distance for this testing data is 2 meters. During experiments, there are 14 scenarios:

1. *Standing position:* in this scenario the person stands up straightly with his/her arms relax. With his/her eyes look at Kinect. The distance between the person's body and Kinect is 176 cm.

2. *Prayer position:* in this scenario the person sits on the ground in a prayer position with his/her right side is in front of the Kinect, the distance between the person's body and Kinect is 150 cm.

3. *Sitting on the floor with stretched legs position:* in this scenario, the person sits on the ground in front of the Kinect with his/her legs are straightened, the distance between the person's body and Kinect is 226 cm.
4. *Sitting on the floor with crossed legs position:* in this scenario, the person sits on the ground in front of the Kinect with his/her legs are crossed, the distance between the person's body and Kinect is 220 cm.

5. *Lying on the floor position:* in this scenario, the person's body lying on the ground in a vertical direction from the viewpoint of the Kinect, the distance between the person's body and Kinect is 126 cm

6. *Lying on the floor with one leg raised position:* in this scenario, the person's body lying on the ground with one leg raised in a vertical direction from the viewpoint of the Kinect, the distance between the person's body and Kinect is 126 cm.

7. *Sitting on the bed position:* the person sits on the bed, with his/her hands are on the bed, and keeps his/her back straight, the distance between the person's body and Kinect is 186 cm.

8. *Standing on the bed position:* in this scenario, the person stands on the bed straightly with his/her arms relax. His/her eyes look at Kinect, the distance between the person's body and Kinect is 226 cm.

9. *Sitting on the bed with stretched legs position:* in this scenario, the person sits on the bed in front of the Kinect with his/her legs are straightened, the distance between the person's body and Kinect is 212 cm.

10. *Lying on the bed with one leg raised position:* in this scenario, the person's body lying on the bed with one leg raised in a vertical direction from the viewpoint of the Kinect, the distance between the person's body and Kinect is 230 cm.

11. *Lying on the bed position:* in this scenario, the person's body lying on the bed in a vertical direction from the viewpoint of the Kinect, the distance between the person's body and Kinect is 230 cm.

12. *Sitting on the chair position:* in this scenario, the person sits on the chair, with his/her hands are on the chair sides, with his/her feet touches the floor, and keeps his/her back straight, the distance between the person's body and Kinect is 221 cm.

13. *Sitting on the chair with the right leg on the left leg position:* in this scenario, the person sits on the chair, with his/her hands are on the chair sides, and his/her right leg on the left leg, and keeps his/her back straight, the distance between the person's body and Kinect is 221 cm.

14. *Sitting on the chair with the left leg on the right leg position:* in this scenario, the person sits on the chair, with his/her hands are on the chair sides, and his/her left leg on the right leg, and keeps his/her back straight, the distance between the person's body and Kinect is 221 cm.

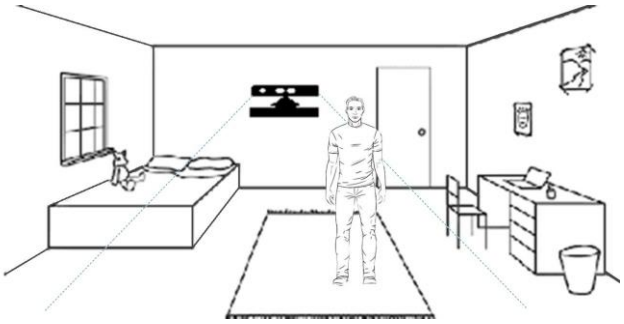

Fig. 7: A simple construction of the room used in our experiments.

V. Results

The SVM model has been trained using our dataset, and the average accuracy is determined by the equation:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

The average accuracy value was 90.2%, then to improve the accuracy value, the random forest model was trained with the best features and get an average accuracy of 99%.

Then the average recall and the precision for each class are also determined by the following equations:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$Precision = \frac{TP}{FP+TP} \quad (5)$$

Where:

TP: represent if the sample is positive and correctly classified as positive, it is considered a true positive.

TN: represent if the sample is negative and correctly classified as negative, it is considered a true negative.

FN: if the sample is positive and classified as negative, it is considered a false negative.

FP: if the sample is negative and classified as positive, it is counted as false positive [16].

TABLE II and TABLE III Show the average recall and the precision for each class for both SVM and random forest models respectively.

TABLE II.
RECALL AND PRECISION VALUES FOR THE SVM MODEL

Class No.	Recall %	Precision %
1.	93.2891	84.8343
2.	88.7235	93.4926
3.	93.2529	91.8025
4.	93.9682	96.4340
5.	89.9020	92.0667
6.	86.9527	86.9803
7.	93.2194	95.0073
8.	92.4460	93.6955
9.	92.7179	94.0658
10.	84.9523	85.4580
11.	84.5845	83.7072
12.	91.7644	93.5132
13.	86.4786	91.1648
14.	90.0427	83.3511

TABLE III.
RECALL AND PRECISION VALUES FOR THE RANDOM FOREST MODEL

Class No.	Recall %	Precision %
1.	99.4529	99.5250
2.	99.8435	99.4483
3.	99.8011	99.8563
4.	99.8140	99.9276
5.	99.8683	99.8135
6.	99.7747	99.7392
7.	99.7773	99.8196
8.	99.8135	99.8135
9.	99.7419	99.8652
10.	99.6608	99.8171
11.	99.7353	99.5520
12.	99.8565	99.8360
13.	99.8368	99.8062
14.	99.7296	99.8376

VI. Conclusion

In this paper, we propose a method for recognizing human activity using the skeleton provided by the Kinect sensor. There are 1120 different experiments with 14 different features extracted from the tracked human skeleton. The obtained results show this skeleton allows classifying well fourteen postures. For classification, the SVM and Random forest techniques are used, from the SVM model get an average accuracy of 92%, and from the Random Forest, get an average accuracy of 99.7%. In the future, we plan to experiment with different classification methods and compare the resulting classification accuracy, also plan to add more experiments to improve the accuracy.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

References

[1] J. Neumann, J. R. Casas, D. Macho, and J. R. Hidalgo, "Integration of audiovisual sensors and technologies in a smart room," *Pers. Ubiquitous Comput.*, vol. 13, no. 1, pp. 15–23, 2009, doi: 10.1007/s00779-007-0172-1.

[2] J. Zhao, G. Zhang, L. Tian, and Y. Q. Chen, "Real-Time Human Detection With Depth Camera Via A Physical Radius-Depth Detector And A Cnn Descriptor School of Computer Science , Shanghai Key Fudan University , China," no. July, pp. 1536–1541, 2017.

[3] S. Majumder and N. Kehtarnavaz, "Vision and Inertial Sensing Fusion for Human Action Recognition: A Review," *IEEE Sens. J.*, vol. 21, no. 3, pp. 2454–2467, 2021, doi: 10.1109/JSEN.2020.3022326.

[4] Microsoft, "Kinect Sensor," no. November, pp. 1371–1372, 2012, doi: 10.13140/2.1.1068.5124.

[5] B. Ben Amor, J. Su, and A. Srivastava, "Action Recognition Using Rate-Invariant Analysis of Skeletal Shape Trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 1–13, 2016, doi: 10.1109/TPAMI.2015.2439257.

[6] A. Jana, "Kinect for Windows SDK Programming Guide," in *PACT Publishing*, 2012, pp. 1–366.

[7] S. Gaglio, G. Lo Re, and M. Morana, "Human Activity Recognition Process Using 3-D Posture Data," *IEEE Trans. Human-Machine Syst.*, vol. 45, no. 5, pp. 586–597, 2015, doi: 10.1109/THMS.2014.2377111.

[8] M. Awad and R. Khanna, "Efficient learning machines: Theories, concepts, and applications for engineers and system designers," *Effic. Learn. Mach. Theor. Concepts, Appl. Eng. Syst. Des.*, no. January, pp. 1–248, 2015, doi: 10.1007/978-1-4302-5990-9.

[9] T. L. Le, M. Q. Nguyen, and T. T. M. Nguyen, "Human posture recognition using human skeleton provided by Kinect," in *2013 International Conference on Computing, Management and Telecommunications, ComManTel 2013*, 2013, pp. 340–345, doi: 10.1109/ComManTel.2013.6482417.

[10] F. Zhu, L. Shao, and M. Lin, "Multi-view action recognition using local similarity random forests and sensor fusion," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 20–24, 2013, doi: 10.1016/j.patrec.2012.04.016.

[11] S. Fallmann and L. Chen, "Computational sleep behavior analysis: A survey," *IEEE Access*, vol. 7, pp. 142421–142440, 2019, doi: 10.1109/ACCESS.2019.2944801.

[12] M. G. A. Komang, M. N. Surya, and A. N. Ratna, "Human activity recognition using skeleton data and support vector machine," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, pp. 0–8, 2019, doi: 10.1088/1742-6596/1192/1/012044.

[13] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A Human Activity Recognition System Using Skeleton Data from RGBD Sensors," *Comput. Intell. Neurosci.*, vol. 2016, 2016, doi: 10.1155/2016/4351435.

[14] A. Ben Tamou, L. Ballihi, and D. Aboutajdine, "Automatic learning of articulated skeletons based on mean of 3D joints for efficient action recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 4, pp. 1–19, 2017, doi: 10.1142/S0218001417500082.

[15] S. Ghazal and U. S. Khan, "Human posture classification using skeleton information," *2018 Int. Conf. Comput. Math. Eng. Technol. Inven. Innov. Integr. Socioecon. Dev. iCoMET 2018 - Proc.*, vol. 2018-Janua, pp. 1–4, 2018, doi: 10.1109/ICOMET.2018.8346407.

[16] L. Wang, D. Q. Huynh, and P. Koniusz, "A Comparative Review of Recent Kinect-Based Action Recognition Algorithms," *IEEE Trans. Image Process.*, vol. 29, pp. 15–28, 2020, doi: 10.1109/TIP.2019.2925285.

[17] Z. Ye and H. Li, "Based on Radial Basis Kernel function of Support Vector Machines for speaker recognition," *2012 5th Int. Congr. Image Signal Process. CISP 2012*, no. Cisp, pp. 1584–1587, 2012, doi: 10.1109/CISP.2012.6469807.

[18] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3542–3549, 2014, doi: 10.1109/CVPR.2014.453.

[19] M. W. Rahman and M. L. Gavrilova, "Kinect gait skeletal joint feature-based person identification," in *Proceedings of 2017 IEEE 16th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2017*, 2017, pp. 423–430, doi: 10.1109/ICCI-CC.2017.8109783.

[20] R. Vemulapalli, F. Arrate, and R. Chellappa, "R3DG features: Relative 3D geometry-based skeletal representations for human action recognition," *Comput. Vis. Image Underst.*, vol. 152, pp. 155–166, 2016, doi: 10.1016/j.cviu.2016.04.005.

[21] M. Alaziz, Z. Jia, R. Howard, X. Lin, and Y. Zhang, "MotionTree: A Tree-Based In-Bed Body Motion Classification System Using Load-Cells," *Proc. - 2017 IEEE 2nd Int. Conf. Connect. Heal. Appl. Syst. Eng. Technol. CHASE 2017*, pp. 127–136, 2017, doi: 10.1109/CHASE.2017.71.

[22] B. Seddik, S. Gazzah, and N. Essoukri Ben Amara, "Human-action recognition using a multi-layered fusion scheme of Kinect modalities," *IET Comput. Vis.*, vol. 11, no. 7, pp. 530–540, 2017, doi: 10.1049/iet-cvi.2016.0326.

[23] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 1010–1019, 2016, doi: 10.1109/CVPR.2016.115.

[24] T. H. An, T. Q. Phuc, N. T. Hai, and T. T. Mai, "Support vector machine algorithm for human fall recognition kinect-based skeletal data," *Proc. 2015 2nd Natl. Found. Sci. Technol. Dev. Conf. Inf. Comput. Sci. NICS 2015*, no. September 2019, pp. 202–207, 2015, doi: 10.1109/NICS.2015.7302191.