Communications for Statistical Applications and Methods
2023, Vol. 30, No. 6, 551–560

https://doi.org/10.29220/CSAM.2023.30.6.551
Print ISSN 2287-7843 / Online ISSN 2383-4757

Motion classification using distributional features of 3D skeleton data

Woohyun Kim[a], Daeun Kim[a], Kyoung Shin Park[b], Sungim Lee[1,a]

[a]Department of Statistics, Dankook University, Korea;
[b]Department of Computer Engineering, Dankook University, Korea

Abstract

Recently, there has been significant research into the recognition of human activities using three-dimensional sequential skeleton data captured by the Kinect depth sensor. Many of these studies employ deep learning models. This study introduces a novel feature selection method for this data and analyzes it using machine learning models. Due to the high-dimensional nature of the original Kinect data, effective feature extraction methods are required to address the classification challenge. In this research, we propose using the first four moments as predictors to represent the distribution of joint sequences and evaluate their effectiveness using two datasets: The exergame dataset, consisting of three activities, and the MSR daily activity dataset, composed of ten activities. The results show that the accuracy of our approach outperforms existing methods on average across different classifiers.

Keywords: motion classification, 3D skeleton data, moments, cross-subject cross-validation, Kinect

1. Introduction

Recognizing and classifying human gestures or motions based on three-dimensional skeleton joint data is computer vision's most active research issue. 3D skeleton-based motion classification is used to build an automatic monitoring system for fall detection or abnormal behavior detection in public places such as hospitals, nursing homes, and large shopping malls (Taha *et al.*, 2015; Shin *et al.*, 2021). Human motion classification is applied in various fields, such as gait studies (Chaaraoui *et al.*, 2015; Tao *et al.*, 2012), healthcare services such as personal training (Lin *et al.*, 2018; Jin *et al.*, 2015), and human-robot interaction in which robots recognize and respond to human actions (Yang *et al.*, 2015; Du *et al.*, 2012). In addition, there is a study focused on monitoring the daily life of the elderly and children in smart home systems (Jalal *et al.*, 2012).

Studies on human motion classification have used data based on wearable sensor devices or computer vision technologies to analyze video images. However, with the recent commercialization of 3D depth cameras such as Kinect, research on utilizing 3D skeleton joint data is increasing. Kinect tracks the human body joints in real-time and provides 3D skeletal information. The Kinect device includes a regular RGB camera and an infrared projector and camera. It consistently calculates the 3D coordinates of various body joints in meters 30 times per second. Kinect is more convenient and cost-effective for collecting 3D skeletal data because no sensors need to be worn, and the computational

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1003257).

[1] Corresponding author: Department of Statistics, 152 Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do 16890, Korea. E-mail: silee@dankook.ac.kr

Published 30 November 2023 / journal homepage: http://csam.or.kr
© 2023 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

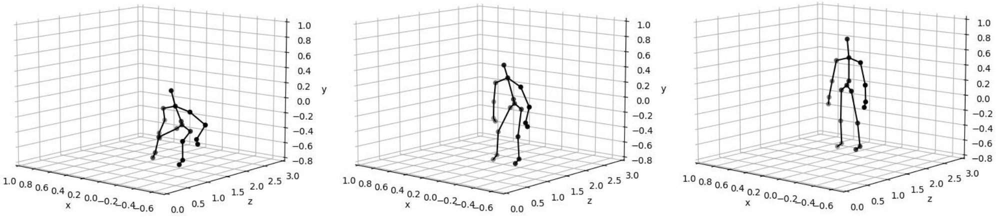

Figure 1: *Example of skeleton joint sequence representing 'stand up' activity, which is from the MSR daily activity dataset (Wang et al., 2012).*

burden is less compared to regular RGB video image processing. The origin of the Kinect system ($x = 0$, $y = 0$, $z = 0$) is positioned at the center of the camera, with the x-axis pointing right, the y-axis pointing upward, and the the z-axis pointing front. The 3D skeleton joint data represents the coordinates of the joints as observed in T-frames. Throughout this paper, the T-frames for a motion are referred to as a skeleton joint sequence. Figure 1 shows a sampled portion of this sequence data of a participant performing 'stand up' from the MSR daily activity dataset (Wang *et al.*, 2012).

In analyzing 3D skeleton joint data for motion classification, it is crucial to consider three key aspects. Firstly, the normalization of the data is an essential step that must be considered. Normalization ensures that all data sequences are set to the same reference point, thus improving the accuracy of the classification process. Failure to normalize the data can lead to the misclassification of motions and activities, even if they are identical, due to differences in the origin of the data. Secondly, the high-dimensional nature of the original 3D skeleton joint data can present a challenge regarding computational burden and performance. Thus, it is necessary to develop and apply functional features that have reduced dimensions to overcome this issue. For example, Park (2016) proposes using new properties, such as the angle and distance between joints, as functional features with reduced dimensions. Finally, it is important to consider the appropriate number of frames to determine the sequence length for motion classification. The number of frames will significantly impact the accuracy and reliability of the classification results. Through this paper, we propose more general features that do not vary according to specific motions or activities for skeleton data and find out how many frames are appropriate for extracting these features.

The paper is organized as follows. Section 2 describes the normalization of the skeleton joint data and general features for representing the joint coordinates. Section 3 presents the moment-based features for classifying 3D skeleton joint data in motion. Section 4 discusses the findings of data analysis of two skeleton-based datasets. Finally, it concludes with a discussion in Section 5.

2. Related work

2.1. Normalization

In Kinect, the j^{th} 3D skeleton joint of a person at frame t is represented as $J_t^j = (x_t^j, y_t^j, z_t^j)$ for $j = 1, 2, \ldots, s$ and $t = 1, 2, \ldots, T$. This 3D skeleton joint data is observed in three dimensions with the number of frames (T) multiplied by the number of joints (s). When we collect this skeleton joint data for a motion or activity through an experiment, it becomes necessary to normalize the data due to different body shapes and movement deviations from different subjects. This pre-processing step is called normalization. Previous studies, such as Cho and Chen (2014), Taha *et al.* (2015), Lee

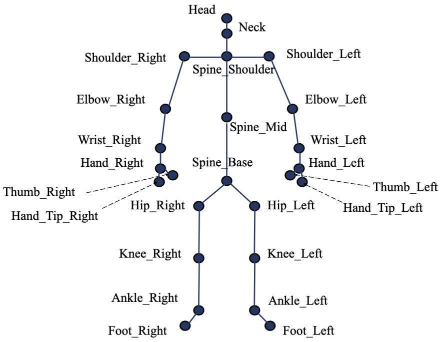

Figure 2: *Skeletal joints of the Kinect scans.*

et al., (2017), Sandra (2020), deal with this pre-processing. In these studies, normalization processing is done by moving the origin of Kinect to the 3D skeleton coordinates so that a particular point on the human body represents the origin $(0, 0, 0)$ in the initial frame. The normalized skeleton joint can be represented mathematically as

$$J^{j}_{t(norm)} = \left(J^{j}_{t} - J^{Center}_{1}\right), \tag{2.1}$$

where $j = 1, 2, \dots, s$ and $t = 1, 2, \dots, T$. Various points on the human body have been used as the origin J^{Center}_{1} for normalization. Lee *et al.* (2017) selected the Spine_Base point in Figure 2 as the reference origin, while Sandra (2020) opted for the Neck point in Figure 2. In contrast, we utilized the midpoint of the spine, denoted as Spine_Med 1 in Figure 2, as the reference origin for normalization. Figure 3 demonstrates the effect of normalization. The left image in Figure 3 shows the 3D coordinates of the initial frame in two sequences for the same subject performing ‘cheer up’ and ‘tossing paper’ from the MSR daily activity dataset. Normalization is crucial in ensuring the consistency of the joint sequences, as it adjusts the starting position of the overall joints. The right image in Figure 3 shows the normalized data, where the initial skeleton coordinates of the two frames are now similar.

2.2. Features in 3D skeleton joint data

In this section, we will analyze the extraction of features when classifying motions from skeleton sequences. There are three commonly used features in skeleton data: (1) a sequence of joint coordinates, (2) temporal differences between consecutive frames, and (3) the mean and range of each joint coordinate. The first feature, joint feature, was used in previous studies such as Patsadu *et al.* (2012), Cho *et al.* (2012), and Du *et al.* (2015). For example, Cho *et al.* (2012) classify arm motions, where the raw joint coordinates are used as predictors. The second feature, motion feature, is the temporal difference of each joint. It is defined as difference of each joint between two consecutive frames:

$$M^{j}_{t} = \left(J^{j}_{t+1} - J^{j}_{t}\right), \quad j = 1, 2, \dots, s, \;\; t = 1, \dots, T - 1. \tag{2.2}$$

Many authors such as Zhu *et al.* (2013), Lee *et al.* (2017), Li *et al.* (2018), and Bagate and Shah (2019) use these motion features to classify various motions. The third feature, summary statistics, was used by Reddy and Chattopadhyay (2014) to differentiate between various movements, such as walking and standing up. Since summary statistics for each joint coordinate are calculated across all the frames, these features reduce the number of predictors compared to joint and motion features. In

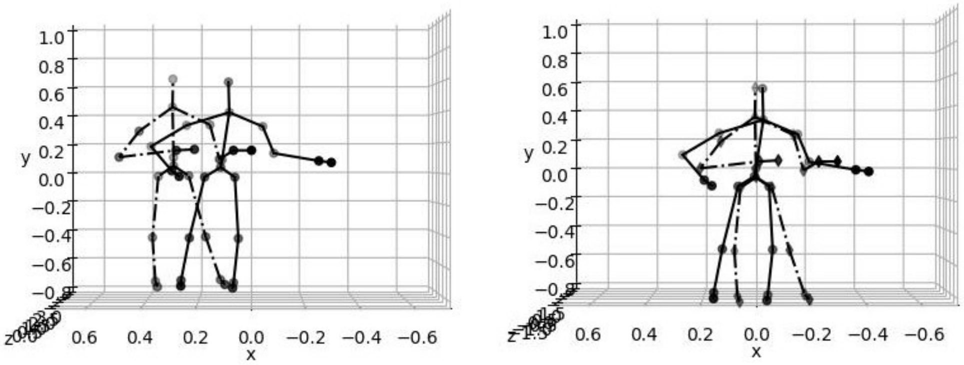

Figure 3: *Different activities performing 'cheer up' and 'tossing paper' from the same subject before normalization (left) and after normalization (right) from the MSR daily activity dataset.*

addition, determining the number of frames corresponds to determining the sample size appropriate for describing the distribution. They used summary statistics with mean and range for j^{th} joint defined as

$$\bar{J}^j = \left(\bar{x}^j, \bar{y}^j, \bar{z}^j\right), \quad j = 1, 2, \ldots, s,$$

$$R\left(J^j\right) = \left(\max_t x_t^j - \min_t x_t^j,\ \max_t y_t^j - \min_t y_t^j,\ \max_t z_t^j - \min_t z_t^j\right),$$

and $\bar{x}^j = (1/T)\sum_{t=1}^{T} x_t^j$. This feature extraction reduces the computational load, but it is uncertain if the mean and range statistics are sufficient to represent the information in the frames. The first and second feature extraction methods have the disadvantage of an increasing number of predictors with an increasing number of frames. The third feature extraction method is simple, but it is questionable if the information in the frames is fully represented. Besides these features, using joint angle targeted by the specific motion or activity (Park, 2016) was proposed, but those features are only general to apply to some of the classification problems.

3. Extracting new distributional features using moments

This paper proposes a novel feature extraction method for motion classification by identifying and extracting the distributional properties of three-dimensional skeleton joint coordinates recorded in T-frames at a rate of 30 frames per second. Instead of utilizing the raw sequence of skeleton joint coordinates, this approach focuses on the distributional aspects of the skeleton joint information over the T-frames. Figure 4 illustrates this concept by showing the distributions of the right-hand joint's x, y, and z-coordinates over 30 frames for walking, running, and tennis motion from Kim *et al.* (2022). This figure demonstrates that the distribution of each coordinate changes based on the motion, even within one second.

Our findings, as illustrated in Figure 4, reveal that the distribution of each coordinate varies according to the type of motion (walking, running, and tennis in Kim *et al.* (2022) experiment). The means, ranges, and shapes of the x, y, and z coordinates differ, leading us to propose a set of summary statistics to describe the distribution of each coordinate. To better characterize a distribution, we can

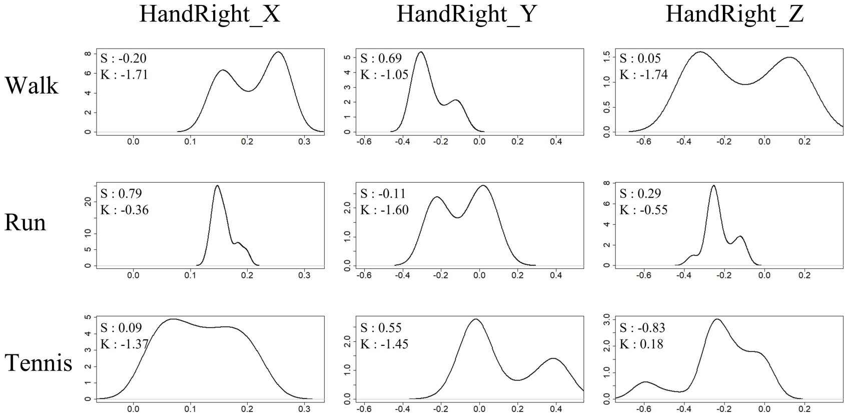

Figure 4: *Density plots of right-hand joint coordinates from 30 frames for one subject according to walk, run, and tennis motion. The S and K stand for skewness and Kurtosis, respectively.*

use four central moment-based statistics: Mean, variance, skewness, and kurtosis. While these quantities may not determine the form of a distribution, higher-order moments are rarely used to describe its characteristics. Mean represents the central location, variance the spread, skewness the asymmetry, and kurtosis the peakedness or shape of the tails. A normal distribution has a skewness of zero, with positive values indicating a right skew and negative values a left skew. If kurtosis is greater than 3, the distribution is leptokurtic with a sharp peak and narrow tails, while a kurtosis less than 3 results in a platykurtic distribution with a flatter peak and wider tails.

The skewness index and kurtosis index are calculated as $\mu_3(\mu_2)^{-3/2}$ and $\mu_4(\mu_2)^{-2}$, respectively, where μ_r is the r^{th} central moment of the random variable X, and μ is its mean. Using the recorded T frames, we estimate the distributional properties (mean, variance, skewness, and kurtosis) from each coordinate of each joint, providing a comprehensive description of the motion.

4. Evaluation

This proposed method is evaluated by using two datasets compared with existing method, such as the linear discriminant analysis (LDA), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), random forest (RF), and gradient boosting machine (GBM) methods for classification with the same settings. The optimal shrinkage parameter in the LASSO classifier was selected through cross-validation, and the number of trees in the RF model was set to 1K. The GBM method was trained with a shrinkage parameter of 0.1 and 1K trees. For the SVM training, the cost value was set to 1, and a gaussian RBF kernel was used. Except for the LDA model, the other four models inherently perform feature selection or regularization, which is useful for high-dimensional data such as skeleton joint data.

When evaluating the models, it is important to consider that some correlations may exist between the sequences when multiple motion sequences are extracted from one person. Therefore, when dividing the samples for cross-validation, sequences extracted from the same subject are placed in the same subset, resulting in a fold equal to the number of subjects.

Table 1: Accuracy (%) comparison in exercise game classification using different features (Joint F, Joint F & Motion F, mean & range, rive number summary, and our proposed four moments) and classifiers (LDA, LASSO, SVM, RF, and GBM) with 30 frames

Features	p	Classifiers					
		LDA	LASSO	SVM	RF	GBM	Avg.
Joint F	2250	85.7	89.0	91.0	88.0	90.5	88.8
Joint F & Motion F	4425	85.0	90.8	92.9	92.3	**96.7**	91.5
Mean&Range	150	94.8	95.6	**94.1**	94.4	95.5	94.9
Five number summary	375	96.9	95.4	92.9	90.4	93.2	93.8
Four moments	300	**97.1**	**96.0**	93.6	**94.7**	95.4	**95.4**

Table 2: Accuracy (%) for classifiers (LDA, LASSO, SVM, RF, and GBM) according to the number of frames using the proposed four-moment features

Classifiers	Number of frames					
	10	20	30	40	50	60
LDA	93.2	95.3	97.1	96.5	96.5	97.0
Lasso	93.1	95.1	96.0	95.0	94.6	94.9
SVM	91.4	93.6	93.6	93.1	93.8	93.5
RF	93.6	94.8	94.7	94.5	94.4	94.3
GBM	92.5	94.2	95.4	95.1	94.9	94.9

Table 3: Confusion matrices obtained through 20-fold cross-subject cross-validation for the proposed method

		Target		
		Walk	Run	Tennis
Prediction	Walk	75.55	2.70	1.15
	Run	0.25	78.70	0.80
	Tennis	1.40	0.70	77.05

Each fold comprises one subject with approximately 80 skeleton sequences in each exercise gameplay.

4.1. Example 1: Classification of exercise games

The dataset by Kim *et al.* (2022) represents 3D skeleton joint data for three exercise gameplays: Walking, running, and playing tennis. Twenty subjects performed each three-minute exercise game using the Nintendo Switch's ring fit adventure game. Kinect captured 5,400 frames for each participant's exercise gameplay. Since the same motion was performed repeatedly in that experiment, we extracted a consecutive sequence of 60 frames, and about 80 sequences of data were collected for each gameplay for each participant. The initial frame of each sequence was adjusted such that the left foot was closest to the ground and there were no overlapping frames. Here, we normalize all sequences' initial frame as in (2.1). The data resulted in a total of 4,766 samples for the three exercise gameplays. The skeleton joint data are high-dimensional, with 25 joints captured in 60 frames, with 2,250 predictors for classification. Hence, appropriate features were necessary for action classification from this dataset.

Five features were used to evaluate the classification performance in this study. They can be classified into two types: One based on raw skeleton joint coordinates and the other based on summary statistics of their distribution. The first type included joint features and joint features combined with motion features, while the second type consisted of mean and range, five summary skeleton joint statistics (Q1, Q2, Q3, minimum value, maximum value), and four moments (mean, variance, skewness, kurtosis), which were proposed in this study.

Table 1 shows the accuracy of various features using LDA, LASSO, SVM, RF, and GBM clas-

Table 4: Summary statistics of the number of frames according to the ten activities from 10 subjects

Frame	Action									
	Cheering	Eating	Lying	Siting down	Siting still	Standing up	Tossing paper	Using cleaner	Walking	Writing
Min	136.0	183.0	103.0	86.0	115.0	77.0	51.0	230.0	67.0	168.0
Avg	173.4	258.7	161.1	121.4	174.2	122.2	106.4	267.8	237.2	231.5
Max	214.0	373.0	206.0	180.0	216.0	182.0	159.0	339.0	359.0	357.0

Table 5: Accuracy (%) comparison in the MSR daily activity 3D dataset using different features (Joint F, Joint F & Motion F, Mean & Range, five number summary, and our proposed four moments) and classifiers (LDA, LASSO, SVM, RF, and GBM) with 30 frames

Features	p	Classifiers					
		LDA	LASSO	SVM	RF	GBM	Avg.
Joint F	1800	81.0	82.0	71.0	85.0	79.0	79.6
Joint F & Motion F	3540	81.0	80.0	62.0	86.0	76.0	77.0
Mean&Range	120	82.0	**91.0**	**89.0**	91.0	87.0	88.0
Five number summary	300	85.0	88.0	87.0	89.0	84.0	86.6
Four moments	240	**93.0**	90.0	**89.0**	**91.0**	**93.0**	**91.2**

Table 6: Confusion matrices obtained through 10-fold cross-subject cross-validations for the proposed method

Prediction	Target									
	Writing	Walking	Using cleaner	Tossing paper	Standing up	Siting still	Siting down	Lying	Eating	Cheering
Writing	0.9	0	0	0	0	0	0	0	0	0
Walking	0	1	0	0	0	0	0	0	0	0
Using cleaner	0	0	0.8	0.1	0	0	0	0	0	0
Tossing paper	0	0	0.1	0.8	0	0	0	0	0	0
Standing up	0	0	0	0	1	0	0	0	0	0
Siting still	0	0	0	0	0	1	0	0	0	0
Siting down	0	0	0	0	0	0	0.8	0	0	0
Lying	0	0	0	0	0	0	0.2	1	0	0
Eating	0.1	0	0.1	0.1	0	0	0	0	1	0
Cheering	0	0	0	0	0	0	0	0	0	1

Each fold comprises one subject performing each of the ten activities once.

sifiers for exercise game classification. It reveals that only joint features result in the lowest performance. Joint features combined with motion features improved the performance of most classifiers, especially GBM. The statistical features perform relatively well regardless of the classifier, with the mean and range features showing 94–95% accuracy. The five summary statistics features perform better with the LDA model, while the moment features have the highest accuracy (97%) with the LDA model and perform relatively well with other classifiers. The GBM method performs similarly using both joints with motion and moment features.

The classifiers utilizing moment features are found to have stable accuracy, with the LDA model outperforming the best. Table 2 shows the classifiers' accuracy for frames between 10 and 60. The results reveal that using more than 30 frames improves the accuracy over 10 or 20 frames in all classifiers. However, the performance is not significantly improved even if the classifiers use more than 30 frames (30, 40, 50, and 60). Therefore, we can infer that sequences of at least ten frames are required, and utilizing distributions of at least 30 frames yields more reliable results. Based on these findings, 30 consecutive frames should be utilized for better classification performance when using moment features.

To identify the source of misclassification, we analyze the results of a 20-fold cross-subject cross-validation (CSCV), which is used to generate the confusion matrix. The confusion matrices are calculated for each fold, and they are averaged over the 20 folds for the LDA model. Note that each fold contains approximately 80 skeleton sequences per gameplay. Table 3 shows that four-moment feature with LDA model effectively classifies the three exercise gameplays. The worst misclassification is ‘run’ classified as ‘walk’ with an average of 2.7 cases. Further examination of the raw data reveals that such a participant is a female, with little difference in her gait between run and walk exercises. Apart from this instance, the LDA model exhibits good classification performance.

4.2. Example 2: Analysis of MSR daily activity data

The MSR daily activity 3D dataset (Wang *et al.*, 2012) is used as a real-world example in this study. The dataset contains 3D skeleton joint data from 10 participants with 16 activities captured by Kinect, where the participant performed each activity twice (once standing and the other sitting). We selected 10 standing activities: Writing, walking, using a vacuum cleaner, tossing paper, standing up, sitting still, sitting down, lying on a sofa, eating, and cheering. The length of the skeleton joint sequences was determined to be $n = 100$. As shown in Table 4, the length of the sequences varied among subjects and activities, ranging from 51 to 553 frames. The activity duration varied by activity type, with ‘using a vacuum cleaner’ being the longest and ‘tossing paper’ having the shortest number of frames. To model the dataset, the number of frames is fixed at 30, and frames are periodically sampled from the entire sequence. The same features and classifiers from section 4.1 are utilized.

Table 5 shows the accuracy of various features using LDA, LASSO, SVM, RF, and GBM classifiers for the MSR Daily Activity 3D dataset. The results show a similar pattern to the previous example. Statistical features improve the model’s accuracy compared to only joint features or a combination of joint and motion features. It is worth noting that the combination of joint and motion features does not contribute to a significant improvement over only joint features, in contrast to the results obtained in the previous dataset analysis. Nevertheless, the LDA and GBM classifiers based on moment features perform the best, with an average accuracy of 93%.

Table 6 shows the results of the misclassification analysis using moments features obtained from a 10-fold CSCV of the LDA model. Each fold in the CV corresponds to one participant performing ten activities. As a result, each cell in the one CV will show either zero or one misclassification. The results are an average of 10 folds. Out of the ten activities, six were accurately classified by all participants, with misclassifications occurring in ‘writing on paper,’ ‘using a vacuum cleaner,’ ‘tossing a paper,’ and ‘sitting down’. Notably, ‘sitting down’ is misclassified as ‘lying on a sofa,’ These two activities were particularly confusing, as they are similar, requiring the participant’s whole body to move downward from a standing up posture to a sitting down posture.

5. Conclusion

This study introduces a novel method for classifying human motion based on 3D skeletal joint data using distributional features. Analysis results of two real examples reveal that the moment-based features outperform traditional features in terms of stability and performance. Moment-based features are versatile because they can be applied regardless of different activities or motions. The proposed approach is based on the premise that joint distributions are distinct in different motions. The two real-data analyses show that these assumptions work well.

Additionally, moment-based features make it easy to determine the appropriate number of frames. This decision is equivalent to determining the number of frames representing the joint coordinate distribution. As a result of evaluating the exercise game dataset, 30 consecutive frames corresponding to one second of information show high classification performance irrespective of classifiers. For the MSR daily activity dataset, 30 frames are sampled at regular intervals in each activity to examine the distribution. The results show that 30 frames perform better using moment-based features than other features, indicating that the joint distribution can be described with that number of frames. However, it is important to note that moment-based features may not be practical when used with a relatively small number of frames which is not enough to represent the distribution. Also, with the MSR daily activity dataset, the distribution of consecutive frames is not practical because each activity has a different frame length, which means that equal numbers of consecutive frames are ineffective at distinguishing between specific activities. As such, successive frames may or may not be practical, depending on the application. Hence we need to sample the frames in a way that works for a given problem. Recent studies have widely used deep learning models for 3D skeleton-based human motion classification, and we will compare these methods with our proposed approach in the future.

References

Bagate A and Shah M (2019). Human activity recognition using rgb-d sensors. In *Proceedings of 2019 International Conference on Intelligent Computing and Control Systems*, Madurai, India, 902–905.

Chaaraoui AA, Padilla-López JR, and Flórez-Revuelta F (2015). Abnormal gait detection with RGB-D devices using joint motion history features. In *Proceedings of 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Ljubljana, Slovenia, 1–6.

Cho B, Jang H, and Zhang B (2012). Motion recognition and classification using kinect sensor data, *KIISE*, **39**, 318–320.

Cho K and Chen X (2014). Classifying and visualizing motion capture sequences using deep neural networks, In *Proceedings of 2014 International Conference on Computer Vision Theory and Applications*, Lisbon, Portugal, 122–130.

Du G, Zhang P, Mai J, and Li Z (2012). Markerless kinect-based hand tracking for robot teleoperation, *International Journal of Advanced Robotic Systems*, **9**, 36, Available from: http://doi:10.5772/50093

Du Y, Wang W, and Wang L (2015). Hierarchical recurrent neural network for skeleton based action recognition, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1110–1118.

Jalal A, Uddin MZ, and Kim TS (2012). Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home, *IEEE Transactions on Consumer Electronics*, **58**, 863–871.

Jin X, Yao Y, Jiang Q, Huang X, Zhang J, Zhang X, and Zhang K (2015). Virtual personal trainer via the kinect sensor. In *Proceedings of 2015 IEEE 16th International Conference on Communication Technology*, conference location, Hangzhou, 406–463.

Kim D, Kim W, and Park KS (2022). Effects of exercise type and gameplay mode on physical activity in exergame, *Electronics*, **11**, 3086, Available from: https://doi.org/10.3390/electronics11193086

Maat S (2020). Clustering gestures using multiple techniques (Doctoral dissertation), Tilburg University, Tilburg, Netherlands.

Lee I, Kim D, Kang S, and Lee S (2017). Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 1012–1020.

Li C, Zhong Q, Xie D, and Pu S (2018). Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, Available from: arXiv preprint arXiv:1804.06055

Lin BS, Wang LY, Hwang YT, Chiang PY, and Chou WJ (2018). Depth camera based system for estimating energy expenditure of physical activities in gyms, *IEEE Journal of Biomedical and Health Informatics*, **23**, 1086–1095.

Park K (2016). Development of kinect-based pose recognition model for exercise game, *KIPS*, **5**, 303–310.

Patsadu O, Nukoolkit C, and Watanapa B (2012). Human gesture recognition using Kinect camera. In *Proceedings of 2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE)*, Bangkok, Thailand, 28–32.

Reddy VR and Chattopadhyay T (2014). Human activity recognition from kinect captured data using stick model. In *Proceedings of International Conference on Human-Computer Interaction*, Heraklion, Crete, Greece, 305–315.

Shin BG, Kim UH, Lee SW, Yang JY, and Kim W (2021). Fall detection based on 2-Stacked Bi-LSTM and human-skeleton keypoints of RGBD camera, *KIPS Transactions on Software and Data Engineering*, **10**, 491–500.

Taha A, Zayed HH, Khalifa ME, and El-Horbaty ESM (2015). Human activity recognition for surveillance applications. In *Proceedings of the 7th International Conference on Information Technology*, Amman, Jordan, 577–586.

Tao W, Liu T, Zheng R, and Feng H (2012). Gait analysis using wearable sensors, *Sensors*, **12**, 2255–2283.

Wang J, Liu Z, Wu Y, and Yuan J (2012). Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 1290–1297.

Yang Y, Yan H, Dehghan M, and Ang MH (2015). Real-time human-robot interaction in complex environment using kinect v2 image recognition. In *Proceedings of 2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics*, Siem Reap, Cambodia, 112–117.

Zhu Y, Chen W, and Guo G (2013). Fusing spatiotemporal features and joints for 3d action recognition, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Portland, OR, USA, 486–491.

Received February 15, 2023; Revised July 06, 2023; Accepted September 14, 2023