ISSN 3047-5473
Innovation in Engineering, Volume 1, Nomor 1, 2024, pp. 1-20

Recognition human walking and running actions using temporal foot-lift features

Khin Cho Tun, Hla Myo Tun*, Lei Lei Yin Win and Khin Kyu Kyu Win

Department of Electronic Engineering, Yangon Technological University, Yangon, MYANMAR

*Corresponding Author: hlamyotun.ytu@gmail.com

DOI: https://doi.org/10.58712/ie.v1i1.1

Abtract: The recognition of human walking and running actions becomes essential part of many different practical applications such as smart video-surveillance, patient and elderly people monitoring, health care as well as human-robot interaction. However, the requirements of a large spatial information and a large number of frames for each recognition phase are still open challenges. Aiming at reducing the number frames and joint information required, temporal foot-lift features were introduced in this study. The temporal foot-lift features and weighted KNN classifier were used to recognize "Walkin and "Running" actions from four different human action datasets. Half of the datasets were trained and the other half of datasets were experimentally tested for performance evaluation. The experimental results were presented and explained with justifications. An overall recognition accuracy of 88.6% was achieved using 5 frames and it was 90.7% when using 7 frames. The performance of proposed method was compared with the performances of existing methods. Skeleton joint information and temporal foot-lift features are promising features for real-time human moving action recognition.

Keywords : Moving action recognition; Skeleton joint data; Foot-lift feature; Smart surveillance system; Human action.

1. Introduction

Today, action-based video retrieval systems, smart video surveillance systems and human-robot interaction require intelligent recognition of human moving actions in a recorded movie or in real-time capturing images. Among different actions in our daily lives, "Walking" and "Running" actions are commonly performed for different activities, especially in public places such as shopping malls, hospitals, university campuses, public parks, banking and ATM (Automatic Teller Machine), relaxation areas, traffic lights as well as condominium residents as shown in Figure 1. In such locations, the seamless observation of people actions is essential to provide a good security. It cannot

Received: January 05, 2024. **Revised**: February 15, 2024. **Accepted**: March 02, 2024

be implemented only by human security and thus the applications of smart video surveillance systems are essentially important [1]. In addition, human recognition is importantly used in human-robot interaction such as human-following robots, sport-robot and so on.

Figure 1: Instantaneous human moving actions in different applications (a) human-following robot (b) public surveillance system (c) sport-robot

In literature, human action recognition was started using the signal of the LED light bulbs attached on different parts of human body [2]. In that approach, the motion vector analysis of LED bulb signal was applied to differentiate the "Walking" and "Running" actions. This approach required dark environment to capture the LED bulb signal. Su et al. [3] proposed an automatic gait recognition method based on the motion frequency characteristics of three parts of human silhouette; trunk, thigh and shin. The "Running" action was marked with a high motion frequency and "Walking" action marked with a low motion frequency. In the approach proposed by Masoud et al. [4] eigen images were used as the features for recognizing "Walk", "Run", "Skip", "March, Line-walk", "Hop", "Side-walk", "Side-skip" actions. It required a total of 50 binary image frames to develop eigen images and eigen vectors. Köse et al. [5] utilized Pixel Motion Features (PMI) for recognition six different actions including "Walking" and "Running".

The PMI feature is the cumulative summation of frame difference for a certain number of consecutive frames during a time duration. The authors used eight cameras to capture eight perspectives of motion and calculate the PMI. The "Running" action resulted in large PMI values and the "Walking" action resulted in small PMI values. The effectiveness of duty factor for recognizing "Walking", "Jogging" and "Running" [6]. Their method consisted of two steps. First, the gait type was classified and then action recognition was recognized based on duty factor in the second step. The duty factor is the total time when one foot is touching to the ground during one stride. In their approach, at least two strides were still needed. These earlier studies have mainly focused on learning and recognizing actions using RGB images. However, application of RGB video/images have limitations when human body parts are occluded due to self-occlusion or objects around when performing action.

Thus, today the application of RGB-D and 3D skeleton-joint information becomes popular for recognition of different human actions . Compared to RGB images, the depth information or skeleton joint data can give 3D posture and additional features of human body during actions [7], [8]. It realizes the possibility of 3D actions recognition [9], [10], [11], [12]. A total of 16 joints features were utilized in the study of Ghazal et al.[13]. Here, the displacements of left/right shoulder, left/right hip, left/right knee left/right ankle between consecutive frames were calculated as dynamic features. Kim et al. [14] proposed weakly-supervised 3D network for recognition of different actions including "Walking" and "Running" actions. However, it still requires at least 16 temporal frames.

From this literature survey, two challenges are observed. The first challenge is related to the requirement of large spatial information. Using skeleton joint information of the whole human body not only increases the computational complexity but also requires a larger memory space because there are 15 to 30 joint locations on entire body [11]. There will be more than 45 data points in each frame when fifteen 3D joints are considered. Thus, when hundreds of frames are used in each recognizing phase, there will be triple number of data to be handled in calculation. Sometimes, multiple cameras or multiple views are necessary for achieving a good classification performance. It not only increases the computational complexity but also requires larger memory space to keep the entire database [15].

The second challenge is the requirement of a large number of frames (in other words long duration of action) for recognizing human action. In [16], 5-s long video with 125 frames were required for each recognition phase. In other works [10], [17], [18], [19] a minimum of 300 frames is required to extract spatiotemporal features in conventional dual-stream model. In approach of using adaptive energy images [20], a total 70 to 100 frames must be applied. Even the whole video clip was used in [21]. These two challenges motivated us to consider and contribute the concept of temporal foot-lift features using a small number of frames for recognizing "Walking" and "Running" actions as suggested in a previous work [22].

In this regard, the objectives of this study are to propose a method for recognizing human "Walking" and "Running" actions using temporal foot-lift features extracted from a small number of frames and to experimentally evaluate the performance using four different datasets.

2. Material and methods

2.1 Skeleton joint data and foot-lift

Today, the skeleton joint data of human body in actions can be extracted using RGB-D camera and Kinet Software [23] . There are a number of skeleton joint information datasets of different human actions and freely available for research purpose [24], [25], [26], [27]. Depending on the number of capturing locations, there can be many key joint

locations from which 3D skeleton joint information are extracted. A sample of skeleton joints for in-door "Walking" action from UTKinet dataset is shown in Figure 2. It has 15 joint locations from head to foot.

This section explains the idea of using foot-lift features. Indeed, people may perform "Walking" and "Running" as a single action or a combined action together with other action. In our daily activities, the upper body part of human can have another activity while walking or running. For examples, carrying box while walking, talking on phone while walking, carrying pole vault or bag while running as shown in Figure 3.

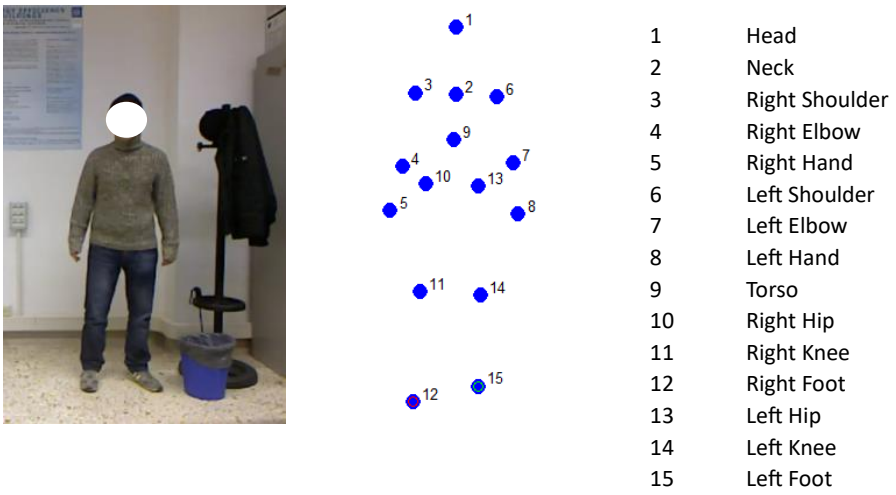


Figure 2: Skeleton joint data [24]

When the whole-body information is taken into account, there are two disadvantages in training the model. First, a large data must be handled and second the upper body part can have many different actions. To avoid these disadvantages and to be able to integrate the proposed method in any activity recognition, only foot-lift is considered for recognizing fundamental "Walking" and "Running" actions. As shown in Figure 4, there are distinct foot-lift characteristics in "Walking" and "Running" actions. It can be noticed that foot-lifts in "Running" action are higher than foot-lifts in "Walking" action. Therefore, it is considered to extract (predictors) features from foot-lift. Then, this concept requires only information of foot joint locations rather than all joints' information. It addresses the first challenge. To address the second challenges, temporal foot-lift features are considered. Instead of using a large number of frames, foot-lift features can be extracted using a small number of frames. In this study, 5~7 frames are taken for each recognition phase.

Figure 3: Some possible human actions while walking and running

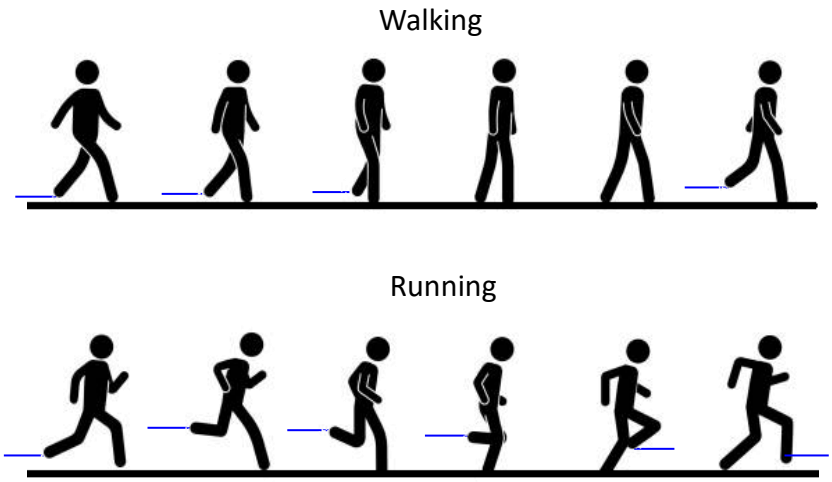


Figure 4: Foot lift in "Walking" and "Running" actions

2.2 Block diagram

Figure 5 shows the process flow diagram of the proposed method. There are two phases in which the first phase is video capturing and skeleton joint data extraction for the actions. The first phase was done by previous research works which developed the datasets [24], [25], [26], [27]. Today, there are many approaches to extract skeleton joint data from recorded images or videos. Some researchers [28] used estimation of vertical position of the neck, shoulder, waist, pelvis, knee, and ankle set by a study of anatomical data to be 0.870H, 0.818H, 0.530H, 0.480H, 0.285H, and 0.039H, respectively. However, it only works for extracting 2D skeleton joints in walking action. The other approach is using various RGB-D sensors such as Microsoft Kinect, Intel RealSense, OrbbecAstraPro and so on. In this approach body parts segmentation is performed in RGB images using readily available deep learning models. Then, the dept information is extracted from depth-image and used to develop 3D posture and skeleton joint data. The other approach is using optical motion capture system based on retro-

reflective markers attached to the actor' s body. This approach was used in MoCap dataset [27].

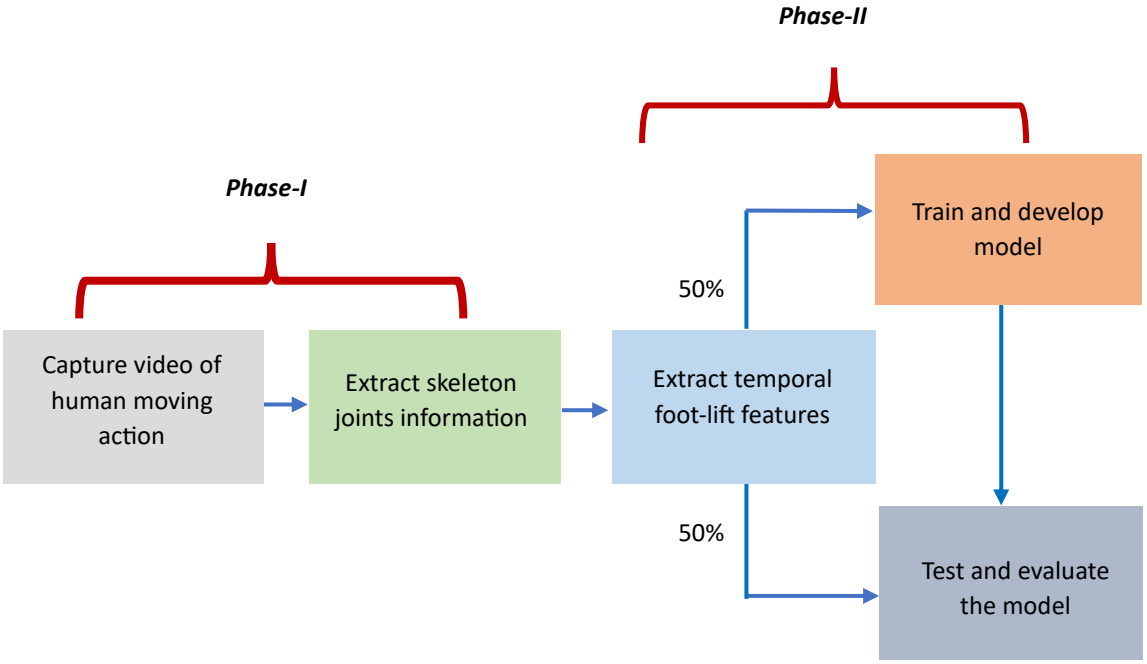


Figure 5. Block diagram of proposed method

The second phase was conducted in this study. Based on the proposed method, the temporal foot-lift features were calculated using skeleton joint information. The detailed procedure for foot-lift feature calculation is explained in the next section. A half of the observations was used for training and developing KNN classification model. Then, the other half of observations was used for testing and evaluating the model.

2.3 Foot-lift features

Before discussing foot-lift features, it is necessary to understand what foot-lift is. The foot-lift is the perpendicular height of the foot from the moving path (e.g, floor, ground) during walking or running as shown in Figure 6. To calculate the foot-lift features, foot-lifts of both feet must be firstly calculated. In this proposed method feet are assigned as lower foot and higher foot but not left and right. It makes calculation more flexible and independent of moving direction.

Figure 6 depicts the concept of calculating foot-lift in an image frame. From a skeleton joint data in an image frame, the coordinates [(x_{fl}, y_{fl}, z_{fl}), (x_{fh}, y_{fh}, z_{fh})] of feet locations were taken. Here, (x_{fl}, y_{fl}, z_{fl}) is lower foot location and (x_{fh}, y_{fh}, z_{fh}) is the higher foot location.

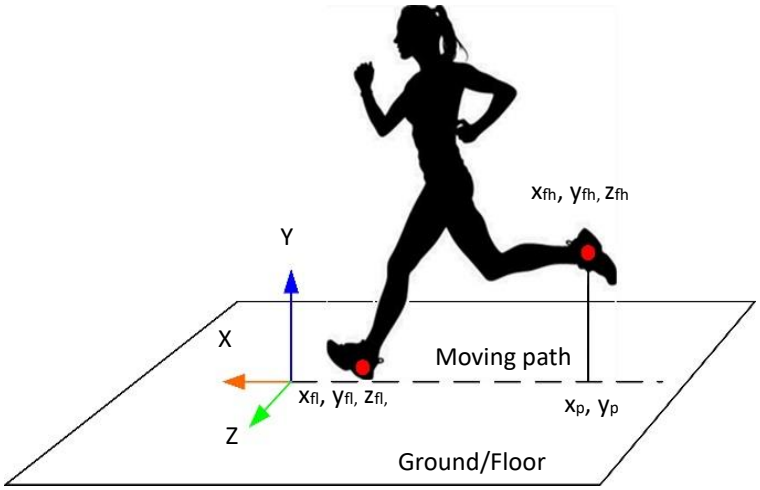


Figure 6: Moving path and foot-lift in a movie frame

First it needs to find a short moving path on the ground/floor. Since it is a short path during a small number of frames, it is considered as a linear line. Thus, straight line equation was used. To find the moving path equation, the coordinate (x_{fl}, y_{fl}, z_{fl}) was used. At least three coordinates are required to develop a linear equation. Therefore, at least three consecutive frames are required to find moving path equation. Here, only 2D (XY or ZY, XY in Figure 6) path line is considered because each recognition phase is implemented during a very short time.

$$y_{fl} = \alpha_1 x_{fl} + \alpha_2 \quad (1)$$

where, $x_{fl}=[x_{fl1}, x_{fl2}, x_{fl3},\cdots, x_{fln}]$, $y_{fl}= [y_{fl1}, y_{fl2}, y_{fl3},\cdots, y_{fln}]$ are vectors of coordinates of lower foot from *n* consecutive video frames. In this work, the number of frames (*n*) is set as 5 and 7. After calculating α_1 and α_2 by means of linear regression method, y_p on the moving path can be calculated at any x_p.

$$y_p = \alpha_1 x_p + \alpha_2 \quad (2)$$

$$\Delta y_h = y_{fh} - y_p \quad (3)$$

$$\Delta y_l = y_{fl} - y_p \quad (4)$$

$$R_{fl} = \frac{\Delta y_l}{H_m} \quad (5)$$

$$R_{fh} = \frac{\Delta y_h}{H_m} \quad (6)$$

where, y_p is y-coordinate on moving path at any foot location x_p (it should be x_{fh} for higher foot location, x_{fl} for higher foot location), Δy_l is lower foot lift, Δy_h is higher foot-lift, H_m is human height, and R_{fl} is normalized lower foot-lift, R_{fh} is normalized higher foot-lift. For example, when *n* is set as 5 or 7, R_{fl} and R_{fh} become vectors as [R_{fl1}, R_{fl2}, R_{fl3}, R_{fl4}, R_{fl5}] and [R_{fh1}, R_{fh2}, R_{fh3}, R_{fh4}, R_{fh5}] respectively.

After calculating foot-lifts for both higher and lower feet, first three foot-lift features were calculated by means of the following equations.

$$\Delta R_{fm} = \bar{R}_{fh} - \bar{R}_{fl} \tag{7}$$

$$R_{fmax} = \max([R_{fh}]) \tag{8}$$

$$E_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R_{fhi} - R_{fli}\right)^2} \tag{9}$$

where, ΔR_{fm} is difference of means of normalized lower foot-lifts and normalized higher foot-lifts, R_{fmax} is maximum normalized foot-lift, E_{rms} is root-mean-square error between lower foot-lifts and normalized higher foot-lifts.

Then, the dynamic features are referred as the ones for which a time duration is required in calculation. In this work, the video recording frame rate is 30 fps. Therefore, the time duration is 0.167 *s* for 5 frames and 0.233 *s* for 7 frames. For this work, two dynamic features were considered as follow. To find the fairness among different physical structures of actors, the normalized values are calculated. These features can be expressed in mathematical forms as follow.

$$PMI_y = \frac{1}{H_m}\sum_{i=1}^{n}\left|y_{fl}(i) - y_{fl}(i+1)\right| + \left|y_{fh}(i) - y_{fh}(i+1)\right| \tag{10}$$

$$PMI_{x,y,z} = \frac{1}{H_m}\sum_{i=1}^{n}\left|x_{fl}, y_{fl}, z_{fl}(i) - x_{fl}, y_{fl}, z_{fl}(i+1)\right| + \left|x_{fh}, y_{fh}, z_{fh}(i) - x_{fh}, y_{fh}, z_{fh}(i+1)\right| \tag{11}$$

where, PMI_y is pixel motion feature in Y-direction and $PMI_{x,y,z}$ pixel motion feature in x, y, z directions. Since only two feet locations are required, there is no effect of the number of joints available. It is a new version of pixel motion feature (PMI). Originally, the PMI can be developed using binary image difference. The larger PMI values represent running action and lower PMI values represent walking action. Here, since RGB images are not used, joints coordinate differences are used. The first one is the cumulative sum of absolute vertical displacements of feet during *n* frames (5, 7 frames or 0.167 *s*, 0.233 *s*). The second one is the cumulative sum of absolute 3D displacements of feet during *n* frames (5, 7 frames or 0.167 *s*, 0.233 *s*).

Sample features for "Running" and "Walking" are shown in Figure 7. It can be seen that typically the foot-lift features of "Running" action are higher than the features of "Walking". It supports in classification. The step-by-step procedure for calculating temporal-foot-lift features is described in Table 1.

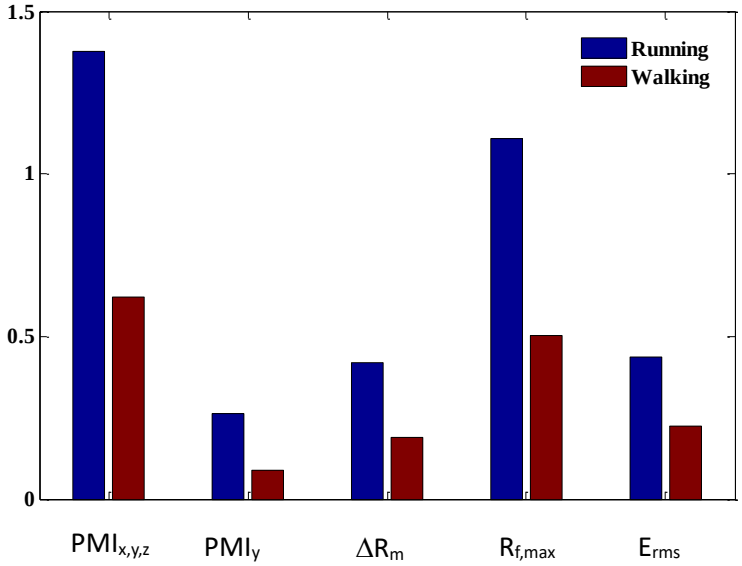


Figure 7: Sample features of "Walking" and "Running" actions

Table 1: Step-by-Step Procedure for temporal-foot-lift feature calculation

Step-by-Step Procedure
- Let the number of frames =5 or 7 or n - Collect lower coordinates of lower foot for 5 frames $x_{fl}=[x_{fl1}, x_{fl2}, x_{fl3},\cdots, x_{fln}]$ and $y_{fl}= [y_{fl1}, y_{fl2}, y_{fl3},\cdots, y_{fln}]$ - Calculate α_1, α_2 using Equation (1) - Calculate y_p using Equation (2) and X-coordinate of higher foot, $x_{fh}=[x_{fh1}, x_{fh2}, x_{fh3},\cdots, x_{fhn}]$ - Calculate higher foot-lifts (Δy_h) using Equation (3) and Y-coordinate of higher foot $y_{fh}=[y_{fh1}, y_{fh2}, y_{fh3},\cdots, y_{fhn}]$ - Calculate lower foot-lifts (Δy_l) using Equation (4) and Y-coordinate of lower foot $y_{fl}=[y_{fl1}, y_{fl2}, y_{fl3},\cdots, y_{fln}]$ - Calculate normalized lower foot-lifts, $R_{fl}=[R_{fl1}, R_{fl2}, R_{fl3}, R_{fl4}, R_{fl5}]$ and higher foot-lifts $R_{fh}= [R_{fh1}, R_{fh2}, R_{fh3}, R_{fh4}, R_{fh5}]$ using Equation (5) and Equation (6) - Calculate foot-lift features, ΔR_{fm}, R_{fmax}, E_{rms} using Equations (7), (8) and (9) - Then, calculate PMI_y and $PMI_{x,y,z}$ using Equations (10), (11)

2.4 Weighted KNN classifier

Today, KNN and SVM classifiers are popularly used in human action recognition. In previous works [16, 24, 25], the Neural Network, KNN, Decision Tree as well as Naïve Bayes were used for classification of human actions. According to the performance comparison in previous studies, KNN have shown its promising accuracy compared to other classifiers. Therefore, weighted KNN with K value of 10 was used in this study. The distance between trained features and test features was calculated using Euclidean distance.

$$d_i = \sqrt{\sum_{j=1}^{m} \left(F_{trained,j} - F_{test,j} \right)^2} \tag{12}$$

$$w_i = \frac{1}{d_i} \tag{13}$$

$$C_{pr} = \frac{\sum_{i=1}^{K} w_i C_i}{\sum_{i=1}^{K} w_i} \tag{14}$$

where, $F_{trained}$ trained feature, F_{test} tested features, m is the number of features (j=[1, 2, 3 ⋯.m], m is 5 in this case), d_i is Euclidean distance for i_{th} observation, w_i is weight factor, C_i is the actual class trained, C_{pr} is the predicted class, and K is the number of nearest distance (K=10 in this study).

2.4 Trained and tested data

For evaluating the performance of proposed method, it was trained and experimentally tested by using four different datasets available in the literature. These datasets are KARD dataset, UTKinet dataset, G3D dataset, CMUMoCap dataset. These datasets were chosen because they include various "Walking" and "Running" actions such as walking in office, walking in building and in-place walking and running.

The number of trained data and test data are shown in Table 2. In KARD dataset and G3D dataset, there are 10 different subjects (persons) who repeat each action 3 times. Then, UTKinet contains dataset for 10 subjects (persons) who repeat each action 2 times. In CMUMoCap dataset, 5 subjects performed "Running" action and 29 subjects performed "Walking" action. However, only "Walking" actions of some subjects were used. "Running" actions are variable only from G3D dataset and CMUMoCap dataset. Both female and male actors are included in performing actions. The

demographic range of actors are 25~40. The participants have given informed consent to use their information in this study.

From each video clip of action, multiple observations can be extracted because only 5 frames or 7 frames were used for each recognition phase. In training the classification model, 50% of each dataset was used and the other 50% of each dataset was used for testing the classification model. Over-fitting can happen when using a large number of observations in training and under-fitting can occur when using a small number of observations in training. Therefore, half-by-half scenario was chosen. In both training and testing, the number of observations for "Walking" action are larger than the number of observations for "Running" action due to availability. The calculations were conducted using MATLAB software.

Table 2: Number of trained and tested observations

Dataset	5 Frames in each observation				7 Frames in each observation			
	Trained data		Test data		Trained data		Test data	
	Run	Walk	Run	Walk	Run	Walk	Run	Walk
KARD	-	48	-	48	-	35	-	34
UTKinet	-	206	-	206	-	147	-	145
G3D	210	213	209	213	150	149	147	155
CMU MoCap	511	498	512	497	357	356	357	355
	721	965	721	964	507	687	504	689
Total	1688		1685		1194		1193	
	3373				2387			

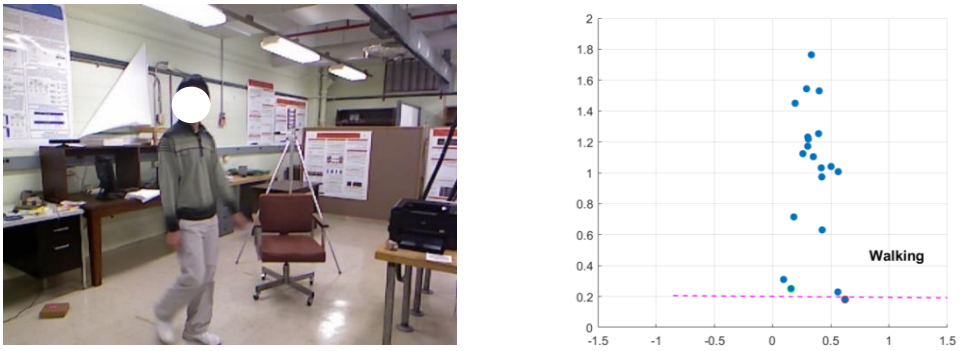
3. Results and discussion

The test results for both "Walking" and "Running" actions from all datasets are shown in Figure 8 (a)-(c) and Figure 9 (a)-(b). In the experimental tests, CMUMoCap dataset has 41 joint locations, G3D dataset has 21 joint locations, UTKinet dataset has 20 joint locations and KARD dataset has 15 joint locations. However, the number of the joints has no effect since only feet joints (two locations) from every dataset were used in this study. The results shown in Figure 8 and Figure 9 are achieved using 7 frames in each recognition phase.

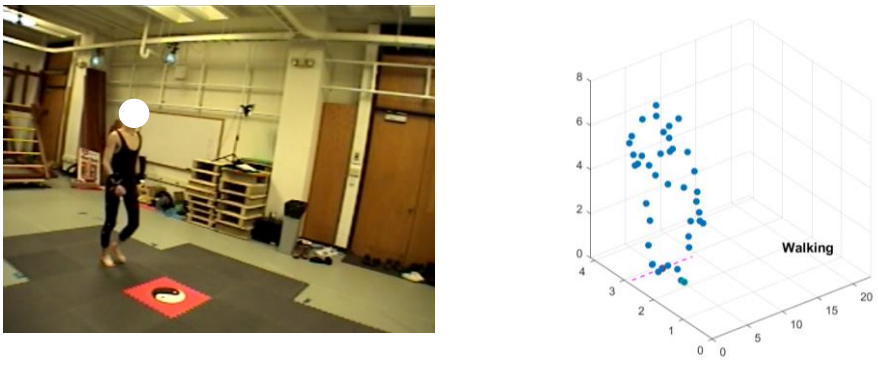
From the results shown in Figure 8 and Figure 9, it can be seen that the subjects performed the actions in different environments and the recordings were conducted from different perspectives. Also, it can be noticed that different subjects have different physical structures and different moving nature. Since the features are normalized using human height, the structure of subjects has no significant effect on the results. The subjects in G3D dataset performed in-place "Walking" and "Running" actions which are

different from normal moving actions. It is like "Walking" and "Running" action in treadmill. The proposed method still works for these conditions since it is based on foot-lift features. The proposed method gives a relatively high accuracy for recognition "Walking" action. However, there are still a few incorrect predictions for "Walking" action. Here, the reasons for incorrect predictions can be explained as follows.

(a)

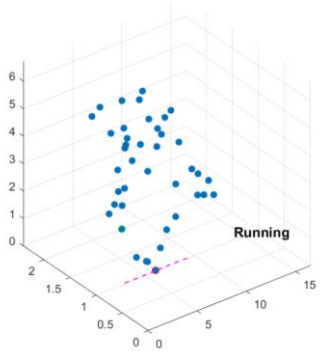


(b)

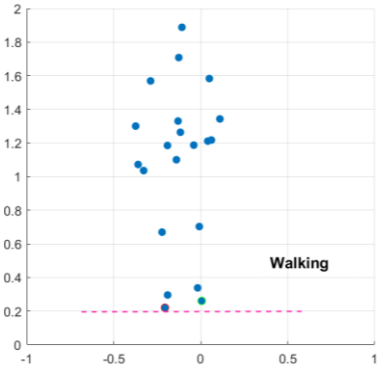


(c)

Figure 8: Recognition results for "Walking" action (a) KARD dataset (b) UT Kinet dataset (c) CMUMoCap dataset



(a)



(b)

Figure 9: Recognition results for "Running" action
(a) G3D dataset (c) CMUMoCap dataset

In running action, there are some moments when both feet are near to ground action as shown in Figure 9(b) and foot-lift features for these conditions are similar to the features of walking action. In addition, the persons performed running action as jogging and the foot-lifts are not obvious and not much different with walking action. There is another reason that when training the video clips running action, there is no obvious foot-lifts in starting frames (start of action) and ending frames (end of action).

On the other hand, in walking action, there are some persons whose foot-lifts are higher than normal due to their physical structure. For those conditions, all foot-lift features are relatively large. Sometimes, in some walking cases, high foot-lifts occur due to inaccurate moving path equation. In some cases, the subject moves around at the end to walk back to original position which results in higher $PMI_{x,y,z}$ value in walking. It leads to some incorrect predictions. One way to improve the accuracy is to increase the number of frames so that observing time is longer and obvious running conditions is captured. The performance of the proposed method can be seen from confusing matrices shown in Table 3 and Table 4. The confusing matrix shown in Table 3 is the performance of

proposed method using 5 frames in each recognition phase. Here, the proposed method gives accuracy of 91.8% for "Walking" action and 85.3% for "Running" action.

Table 4 shows the confusing matrix using 7 frames in each recognition phase. It can be seen that the performance of proposed method is increased for both "Walking" and "Running" actions, 92.1% and 89.3% respectively. The proposed method can give a high performance using only 7 frames instead of using hundred frames or the whole action video clip for each recognition phase. Thus, it is potential for real-time human-robot interaction which needs instantaneous and continuous recognition.

Table 3: Confusion matrix using 5 frames

	Walking	Running
Walking	885 (91.8%)	79 (8.1%)
Running	106 (14.7%)	615 (85.3%)

Table 4: Confusion matrix using 7 frames

	Walking	Running
Walking	635 (92.1%)	54 (8.6%)
Running	54 (10.7%)	450 (89.3 %)

Table 5 shows the recognition accuracy achieved for each dataset. Here, it can be seen that the proposed method is most fitted with MoCap dataset because it has both "Running" and "Walking" actions. Also, the feature, $PMI_{x,y,z}$, is obvious for moving around action. Here, it can be noticed that the proposed method shows a better performance for "Running" action in MoCap dataset than for "Running" action in G3D dataset. Since "Running" action in G3D dataset is in-place running, it results low $PMI_{x,y,z}$ value compared to "Running" action in MoCap dataset. For "Walking" action, the proposed method shows a better accuracy for KARD dataset compared to UTKinet dataset.

Table 5: Recognition accuracies for different datasets

Number of frames	Overall Accuracy %	UTKinet Dataset Test Accuracy % (Walking)	KARD Dataset Test Accuracy % (Walking)	G3D Dataset Test Accuracy % (Walking)	G3D Dataset Test Accuracy % (Running)	MoCap Dataset Test Accuracy % (Walking)	MoCap Dataset Test Accuracy % (Running)
5	88.6	90.1	91.5	92.5	81.3	93.1	89.4
7	90.7	90.2	92.1	92.7	85.3	93.5	93.4

Table 6: Comparison with state-of-the-art-methods

Method	Dataset used	Joint information used	Number of frames used	Action	Accuracy (%)
[6]	KTH Datasets	The whole silhouette	60	Walking, Jogging Running	92.0
[22]	Weizmann, KTH Datasets	The whole silhouette	7	Walking and Running	98.5
[22]	Weizmann, KTH Datasets	The whole silhouette	10	Walking and Running	99.6
[29]	KTH Datasets	The whole silhouette	The whole video clip	Walking and Running	93.8
[30]	Weizmann Datasets	The whole silhouette	10	Walking and Running	99.6
[16]	NTU RGB+D dataset	12, 15, 39	125	Walking	100.0
Ours	KARD, UTKinet, G3D, MoCap	2	7	Walking and Running	90.7

Table 6 shows the performance of proposed method comparing with existing methods. Although most existing methods showed impressive accuracies, they still have challenges in joint information and frame requirements. It can be seen that proposed methods in [31], [22], [29], [30] were based on RGB images and the information of the whole silhouette. Thus, the existing method could result in a high uncertainty in occluded conditions. Meanwhile the proposed method in this study requires only two (feet) joint information.

In other works [31], [29] [16] the number of frames required is 60, 125 or the whole video clip. In [22] and [30] the number of frames required is 7 to 10 to obtain an accuracy of up to 99.6%. However, it was based on the whole silhouette information. Also, the tested data are from only two datasets. Although the proposed method showed a little lower than existing methods, requirements of only feet information and very small number of frames are preferable points of proposed method. It is very potential to be utilized in real-field applications. The other advantage of the proposed method is that it still works with in-place walking or running which are rarely considered in previous

studies. Therefore, the proposed methods work well with a wide range of "Walking" and "Running" action.

4. Conclusion

In this study, temporal foot-lift features were used to recognize "Walking" and "Running" actions aiming at reducing the number of frames required for action recognition. The skeleton joints data from four popular datasets (KARD, UTKinet, G3D, CMUMoCap datasets) were used for both training and testing classification model. The weighted KNN classification method was used. The concept of using temporal foot-lift feature was introduced. The temporal foot-lift features were calculated by using only two foot joint information. The number of frames used in each recognition phase is 5 and 7. The results showed that the proposed method can give accuracy of 91.8% for "Walking" action 85.3% for "Running" action using 5 frames in each recognition phase. When 7 frames were used, the performance of proposed method was increased to 92.1% and 89.3% respectively.

Temporal foot-lift features are very useful for recognizing fundamental moving actions without needing the upper body part information. Thus, foot-lifts features can be used in recognition of in-door/out-door daily-life activities and sport activities by combining with other additional features. The current study has some limitations. There are mixing moments of "Running" and "Walking" which cause incorrect recognition.

In future works, the more performance analysis should be done using a a larger number of frames. On the other hand, frame dropping strategy should be considered to remove similar frames. Thus, the optimum number of frames can be found in further works. Also, jogging and standing actions should be taken into account. Then, more features should be considered in future works.

Author contribution

Khin Cho Tun implemented coding, did experimental tests and prepared the manuscript. Hla Myo Tun developed the main idea of foot-lift features, designed the research, supervised the research work, and guided the preparation of the research paper. Lei Lei Yin Win helped in final proof-reading of the manuscript. Khin Kyu Kyu Win help in finding data source and proof-reading.

Funding statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements

The authors would like to express their sincere thanks to all research partners from both who have developed and shared valuable datasets for the further studies including current study related to human action recognition.

Competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Kashef, A. Visvizi, and O. Troisi, "Smart city as a smart service system: Human-computer interaction and smart city surveillance systems," *Comput Human Behav*, vol. 124, pp. 1–14, Nov. 2021, https://doi.org/10.1016/j.chb.2021.106923

[2] G. Johansson, "Visual perception of biological motion and a model for its analysis"," *Percept Psychophys*, vol. 14, no. 2, pp. 201–211, Jun. 1973.

[3] H. Su and F.-G. Huang, "Human Gait Recognition Based on Motion Analysis," in *International Conference on Machine Learning and Cybernetics*, Guangzhou, China: IEEE, Aug. 2005, pp. 4464–4468. https://doi.org/10.1109/ICMLC.2005.1527725

[4] O. Masoud and N. Papanikolopoulos, "A method for human action recognition," *Image Vis Comput*, vol. 21, no. 8, pp. 729–743, Aug. 2003, https://doi.org/10.1016/S0262-8856(03)00068-4

[5] N. Käse, M. Babaee, and G. Rigoll, "Multi-view human activity recognition using motion frequency," in *IEEE International Conference on Image Processing (ICIP)*, Beijing, China: IEEE, Sep. 2017, pp. 3963–3967. https://doi.org/10.1109/ICIP.2017.8297026

[6] P. Fihl and T. B., "Recognizing Human Gait Types," *Robot Vision*, pp. 183–208, Mar. 2010, https://doi.org/10.5772/9293

[7] T. Ahmad, S. T. H. Rizvi, and N. Kanwal, "Transforming spatio-temporal self-attention using action embedding for skeleton-based action recognition," *J Vis Commun Image Represent*, vol. 95, pp. 1–11, Sep. 2023, https://doi.org/10.1016/j.jvcir.2023.103892

[8] Y. Hbali, S. Hbali, L. Ballihi, and M. Sadgal, "Skeleton-based human activity recognition for elderly monitoring systems," *IET Computer Vision*, vol. 12, no. 1, pp. 16–26, Feb. 2018, https://doi.org/10.1049/iet-cvi.2017.0062

[9] X. Jiang, K. Xu, and T. Sun, "Action Recognition Scheme Based on Skeleton Representation with DS-LSTM Network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2129–2140, Jul. 2020, https://doi.org/10.1109/TCSVT.2019.2914137

[10] A. F. Bavil, H. Damirchi, and H. D. Taghirad, "Action Capsules: Human Skeleton Action Recognition," *Computer Vision and Image Understanding*, vol. 223, pp. 1–11, Aug. 2023, https://doi.org/10.1016/j.cviu.2023.103722

[11] M. A. R. Ahad, M. Ahmed, A. Das Antar, Y. Makihara, and Y. Yagi, "Action recognition using kinematics posture feature on 3D skeleton joint locations," *Pattern Recognit Lett*, vol. 145, pp. 216–224, May 2021, https://doi.org/10.1016/j.patrec.2021.02.013

[12] F. Khezerlou, A. Baradarani, and M. A. Balafar, "A Convolutional Autoencoder Model with Weighted Multi-Scale Attention Modules for 3D Skeleton-Based Action Recognition," *J Vis Commun Image Represent*, vol. 92, pp. 1–14, Apr. 2022.

[13] S. Ghazal, U. S. Khan, M. M. Saleem, N. Rashid, and J. Iqbal, "Human activity recognition using 2D skeleton data and supervised machine learning," *IET Image Process*, vol. 13, no. 13, pp. 2572–2578, Nov. 2019, https://doi.org/10.1049/iet-ipr.2019.0030

[14] J. Kim, G. Li, I. Yun, C. Jung, and J. Kim, "Weakly-supervised temporal attention 3D network for human action recognition," *Pattern Recognit*, vol. 119, pp. 1–10, Nov. 2021, https://doi.org/10.1016/j.patcog.2021.108068

[15] W. Peng, X. Hong, and G. Zhao, "Tripool: Graph triplet pooling for 3D skeleton-based action recognition," *Pattern Recognit*, vol. 115, pp. 1–12, Jul. 2021, https://doi.org/10.1016/j.patcog.2021.107921

[16] M. Terreran, L. Barcellona, and S. Ghidoni, "A general skeleton-based action and gesture recognition framework for human–robot collaboration," *Rob Auton Syst*, vol. 170, pp. 1–14, Dec. 2023, https://doi.org/10.1016/j.robot.2023.104523

[17] Q. Xu, W. Zheng, Y. Song, C. Zhang, X. Yuan, and Y. Li, "Scene image and human skeleton-based dual-stream human action recognition," *Pattern Recognit Lett*, vol. 148, pp. 136–145, Aug. 2021, https://doi.org/10.1016/j.patrec.2021.06.003

[18] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Computer Vision and Image Understanding*, vol. 208–209, pp. 1–10, Jul. 2021, https://doi.org/10.1016/j.cviu.2021.103219

[19] Q. Ye, Z. Tan, and Y. Zhang, "Human action recognition method based on Motion Excitation and Temporal Aggregation module," *Heliyon*, vol. 8, no. 11, pp. 1–12, Nov. 2022, https://doi.org/10.1016/j.heliyon.2022.e11401

[20] O. C. Kurban, N. Calik, and T. Yildirim, "Human and action recognition using adaptive energy images," *Pattern Recognit*, vol. 127, pp. 1–23, Jul. 2022, https://doi.org/10.1016/j.patcog.2022.108621

[21] J. Lin, Z. Mu, T. Zhao, H. Zhang, X. Yang, and P. Zhao, "Action density based frame sampling for human action recognition in videos," *J Vis Commun Image Represent*, vol. 90, pp. 1–7, Feb. 2023, https://doi.org/10.1016/j.jvcir.2022.103740

[22] K. Schindler, E. Zürich, L. Van Gool, and K. Leuven, "Action Snippets: How many frames does human action recognition require?," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA: IEEE, Jun. 2008, pp. 1–8. doi: https://doi.org/10.1109/CVPR.2008.4587730

[23] M. B. Shaikh and D. Chai, "Rgb-d data-based action recognition: A review," *Sensors*, vol. 21, no. 12, pp. 1–25, Jun. 2021, https://doi.org/10.3390/s21124246

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

[25] V. Bloom, V. Argyriou, and D. Makris, "Hierarchical Transfer Learning for Online Recognition of Compound Actions," *Computer Vision and Image Understanding*, vol. 144, pp. 62–72, Mar. 2015, https://doi.org/10.1016/j.cviu.2015.12.001

[26] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal, "View Invariant Human Action Recognition Using Histograms of 3D Joints ," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, RI: IEEE, Jul. 2012, pp. 20–27. https://doi.org/10.1109/CVPRW.2012.6239233

[27] re3data.org, "CMU Graphics Lab Motion Capture Database," re3data.org - Registry of Research Data Repositories. Accessed: Apr. 01, 2024. [Online]. Available: http://mocap.cs.cmu.edu/

[28] J. H. Yoo and M. S. Nixon, "Automated markerless analysis of human gait motion for recognition and classification," *ETRI Journal*, vol. 33, no. 2, pp. 259–266, Apr. 2011, https://doi.org/10.4218/etrij.11.1510.0068

[29] H. Jhuang, T. Serre, and L. Wolf, "A Biologically Inspired System for Action Recognition," in *IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil: IEEE, Oct. 2007, pp. 1–8.

[30] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *IEEE Trans Pattern Anal Mach Intell*, vol. 29, no. 12, pp. 1395–1402, Dec. 2007, https://doi.org/10.1109/TPAMI.2007.70711

[31] P. Fihl and T. B., "Recognizing Human Gait Types," in *Robot Vision*, InTech, 2010. https://doi.org/ 10.5772/9293

Nomenclature

C	Class
d	Euclidean distance
E	Error (Difference)
F	Feature
K	Number of nearest neighbours
n	Number of frames
H	Height of human
PMI	Pixel motion feature
R	Normalized foot-lift
w	Weight
x	X-coordinate
y	Y-coordinate
z	Z-coordinate

Subscripts

f	foot
h	higher
i, j	i^{th} , j^{th}
l	lower
p	moving path,
pr	predicted
m	mean
max	maximum
rms	root-mean-square
trained	trained
tested	tested
x	x-direction

y y-direction
z z-direction

Symbols

α Coefficeints
Δ Difference