Engineering Journal issn: 0125-8281

Article

Human Motion Recognition Using Temporal Foot-Lift Features Extracted from a Small Number of Skeleton Data Frames and Multi Classifiers

Khin Cho Tun[a], Hla Myo Tun[b,*], and Khin Kyu Kyu Win[c]

Department of Electronic Engineering, Yangon Technological University, Myanmar
E-mail: [a]khinchotun@gmail.com, [b,*]hlamyotun.ytu@gmail.com (Corresponding author)
[c]khinkyukyuwin.ygn@gmail.com

Abstract. Human motion recognition becomes an essential part of human–robot collaboration in many different applications such as robot-assisted smart factories, smart warehouse and smart transportation. However, there are still challenges in terms of spatial information and temporal information requirements. Aiming at reducing the number of frames and joint information required, temporal foot-lift features were introduced in this study. The temporal foot-lift features and five different classifiers were applied to recognize "Walking" and "Running" actions from four different human action datasets. Half of the data were trained and the rest were experimentally tested for performance evaluation. The results revealed that the proposed method can give up to 100% accuracy even using a small number of frames. Using KNN classifier and temporal foot-lift features can give the highest performance in recognition. The performance of proposed method was compared with existing methods' performance. The skeleton joint information and temporal foot-lift features are promising features for real-time human motion action recognition.

Keywords: Human-robot interaction, human motion recognition, skeleton joint data, foot-lift feature, smart surveillance system, KNN classification.

ENGINEERING JOURNAL Volume 28 Issue 7
Received 10 April 2024
Accepted 17 July 2024
Published 31 July 2024
Online at https://engj.org/
DOI:10.4186/ej.2024.28.7.41

DOI:10.4186/ej.2024.28.7.41

1. Introduction

Today effective human–robot collaboration becomes essential in many different applications such as robot-assisted smart factories, smart warehouse, smart synergy lab, smart transportation and so on. In human-robot collaboration, the robot must be able to recognize the human motion in order to prevent any potential accident and to ensure the safety condition at workplace [1]. Today, many researchers have been working on advanced and innovative approaches for human motion recognition. Some decades ago, human action recognition was started basing on simple silhouette detection in RGB video frames, then towards using RGB-D frames. Currently, more innovative approaches have been developed using the power of deep learning to recognize action in RGB video frames or RGB-D frames [2, 3, 4, 5]. In parallel, the application of skeleton joint data has gained attraction in human action recognition [6, 7, 8].

Since human actions are composed of consecutive movements of body parts over time, the extraction of spatial-temporal features is the most important part of human action recognition algorithms [9, 10, 11]. Generally, the main differences among previous approaches are taking a short duration or long duration or even the whole clip of action and a portion or the whole-body joint data. Although most methods have achieved impressive results, the requirement of a large number of frames (in other words long duration of action) is still one open challenge. It is related to temporal features. In [1], 5-s long video with 125 frames were required for each recognition phase. In other works [5, 6, 12, 13], a minimum of 300 frames was required to extract spatiotemporal features in conventional dual-stream model. In approach of using adaptive energy images [14], a total 70 to 100 frames must be applied. Even the whole video clip was used in [15]. This challenge has been considered in [16] and the authors worked on it using a small number of frames.

The other challenge is the requirement of skeleton joint data of the whole body. It is related to spatial features. Since there are 15 to 30 important joint locations on entire human body [17], there will be at least 45 data points in each frame. When hundreds of frames, sometimes the whole video clip, were taken there will be triple number of data to be handled in processing. It increases not only the computational complexity but also computational cost as well as required space [18]. The challenges are illustrated in Fig.1.

Although it is reasonable to use the whole video clip or a large number of frames for short/impulse actions to capture a complete action, it is not worth for sinusoidal human motion such as "Walking" and "Running" actions. Human motion actions, especially "Walking" and "Running" actions can be recognized only based on spatial transformation of foot joint without depending upper body joints data. Therefore, this study is to address both challenges occurred in spatial and temporal perspectives. For this reason, an efficient recognition framework for human motion recognition using temporal foot-lift features extracted from a small number of skeleton data frames is proposed in this study. The foot-lift features were extracted from a smaller number of frames compared to existing works.

The rest of this paper is structured as follows: Section 2 introduces some related works. In Section 3, temporal foot-lift features based human motion recognition framework is described. Section 4 provides the experimental results and discussion. Finally, conclusion remarks are given in Section 5.

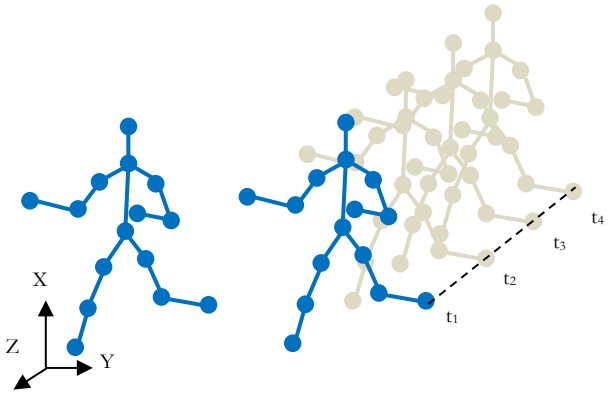


Fig. 1. (a) Spatial information in a frame (b) Temporal information of joint data.

2. Related Works

In literature, most works are forwarding to increasing the recognition accuracies by means of innovative approaches by relying on high-speed processing devices today. However, a high complexity and handling a massive data in training and computation can bring difficulties in real-time applications due to long response time and limited storage resource capacity of edge devices [8]. For this reason, there are some works which considered simplicity, light-weight and real-time processing or response time while maintaining the comparative performance. Knowledge distillation is one trending approach to reduce the weight of the deep learning model [8, 18, 19, 20].

In a recent work [16], the authors tried to answer a question "how many frames are required to perform action recognition?". The authors worked with very short snippets (1-7 frames) in order to find the performance of their proposed method which is based on RGB image frames, template features and SVM classification method. The authors worked with actions in KTH dataset and Weizmann dataset for performance evaluation. The authors concluded that the basic actions can be recognized well using even very short snippets of 1-7 frames (at the frame rate of 25 fps). This study inspired us to work on reducing the number of frames required in action recognition.

Sometimes, spatial-temporal occlusion occurs in skeleton joint data. Here, temporal occlusion is missing

frame in sequence. This type of problem was addressed in [21] by proposing multi-view information fusion method. This work also motivated us to consider only foot-lift features by assuming the upper body part joint information are not available.

In order to save computational effort, skipping the redundant information in video sequence is also as important as finding missing information. Optical flow feature is commonly used for human action recognition but it consumes time due to necessity of computation at every consecutive frame [22]. For this reason, the authors proposed dynamic frame skipping concept by inspecting similarity of two consecutive frames. The same concept was proposed in [23]. The authors considered frame scrapping by checking action difference in the conceptive frames. In another work [24], key-frame sampling method was introduced to reduce redundant frames. The authors of one previous work [25] proposed active stream monitoring mask to extract only active frames which contain certain action event. It can remove both unnecessary regions in the images and unnecessary frames, which consequently reduce computational overheads and the process becomes a light-weight one. Then, key frames extraction method was applied in recent works [26, 27]. This literature review highlights the effort of scholars to reduce the computational overhead and to realize real-time implementation. The current research work is a piece of contribution to lightweight algorithm development.

3. Proposed Recognition Method

3.1. Recognition Phase

In literature, there are two approaches of human action recognition; in the first approach, the features are extracted using all frames form the whole video clip and action label is assigned to the entire video clip as shown in Fig. 2 (a). In the second approach, temporal features are extracted from as small set of frames and assigns the action label to each group [16] as shown in Fig. 2 (b). The first approach is only feasible for the whole recorded video clip which contains only one completed action. Also, the first approach is suitable for recognition in video retrieval system which searches and extracts videos which contain the desired action from recorded database. Using the first approach, it is not possible to recognize action instantaneously during the action. In addition, it will not work if there are different actions or action change in the video clip.

The second approach overcomes this issue and thus it can be realized in practical implementation. The second approach can be implemented in real-time action recognition systems for long and sinusoidal actions or a series of different actions. Therefore, the current work is based on the second approach using 5, 7, 9, 15, 20 frames in each recognition phase.

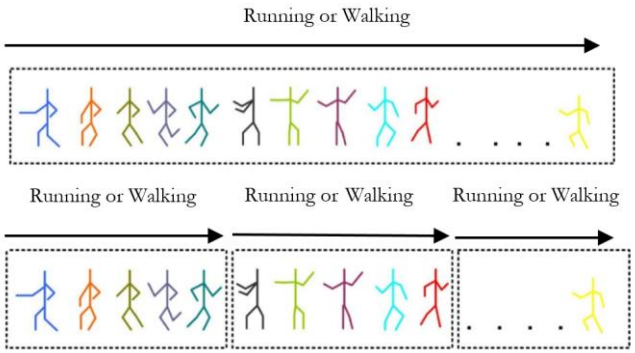


Fig. 2. (a) Recognition phase based on the whole video clip (b) Recognition based on temporal video frames.

3.2. Process Flow

The process flow diagram of the proposed recognition framework is shown in Fig. 3. The first step is capturing RGB-D video images of the action. It can be a recorded video or real time capturing video. The second step is the extraction of skeleton joint information from the video frames. In literature, many different methods have been utilized for getting skeleton joint data from recorded images or videos. In some works, the position of the neck, shoulder, waist, pelvis, knee, and ankle were located using ratio values of 0.870, 0.818, 0.530, 0.480, 0.285, and 0.039 respectively with respective to human height. Today, RGB-D sensors such as Microsoft Kinect, Intel RealSense, OrbbecAstraPro are also popularly used to extract skeleton joint data. In another approach used in MoCap dataset, retro-reflective markers were attached to the actor's body.

In the third step, the temporal foot-lift features are calculated using skeleton joint information. Finally, the features are input to the multi classification models to classify the action. Here, training process to develop classification models and testing process to evaluate the models happen in the same procedure. A half of the observations was used for developing classification models and the other half of observations was used for testing and evaluating the models.

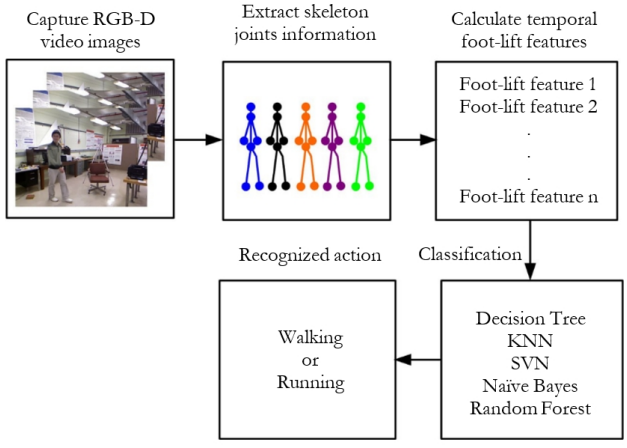


Fig. 3. Process flow diagram.

3.3. Input Dataset

Four different skeleton joint datasets were used in this study. These datasets are KARD dataset [28], UTKinet dataset [29], G3D dataset [30], CMUMoCap dataset [31] which are readily available in literature for research purpose and have been used in many of research studies.

KARD dataset: consists of datasets of 18 activities performed by 10 different persons. Every person performed 3 times for each action. From many different activities, data for walking action were extracted. The dataset provides four types of data; depth map; 640 × 480 RGB video; skeleton joint data in real world coordinates; skeleton joint data on screen coordinates. The action videos are captured at 30 fps.

UTKinet dataset: The video frames were captured using a single stationary Kinect camera. It includes different action images of 10 individuals: walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands as well as clap hands. For these actions, three different data types are available. These are RGB images, depth images and skeleton joint data. The images were recorded in 30 fps. Thus, this dataset has data in the form of a series of images with resolution of 480 × 640.

G3D dataset: This dataset contains 20 gaming actions including "Walking" and "Running". There are 10 actors for these actions. Each actor performed three times for each action. In this dataset, depth image, RGB image (640 × 480 PNG) and skeleton joint data are available. The images are taken at 30 fps.

CMUMoCap dataset: In CMUMoCap dataset, five subjects performed "Running" action and 29 subjects performed "Walking" action. However, only "Walking" actions of some subjects were used. The "Running" actions are variable only from G3D dataset and CMUMoCap dataset.

3.4. Foot-Lift Features

First, it is necessary to understand what foot-lift is and why it is chosen as features. The foot-lift is the perpendicular height of the foot from the moving path (e.g, floor or ground) during walking or running as shown in Fig. 4. In practical human "Walking" and "Running" motions, there is a distinct foot-lift characteristics. It is obvious that the foot-lifts in "Running" action are higher than the foot-lifts in "Walking" action. Therefore, it is considered to extract prediction features from foot-lift. Then, the proposed concept does not require the information of the whole body which may be partially occluded sometimes. Then, temporal foot-lift features can be extracted using a small number of frames according to the aim of this study.

To calculate foot-lift features, foot-lifts of both feet must be firstly calculated. One advantage of this proposed method that feet are assigned as lower foot and higher foot but not left and right. It makes calculation more flexible and independent of starting point and moving direction. Figure 4 depicts the concept of calculating foot-lift in a movie frame. From a skeleton joint data in a movie frame, the coordinates [(x_{f1}, y_{f1}, z_{f1}), (x_{fh}, y_{fh}, z_{fh})] of feet locations were taken. Here, (x_{f1}, y_{f1}, z_{f1}) is the lower foot location and (x_{fh}, y_{fh}, z_{fh}) is the higher foot location.

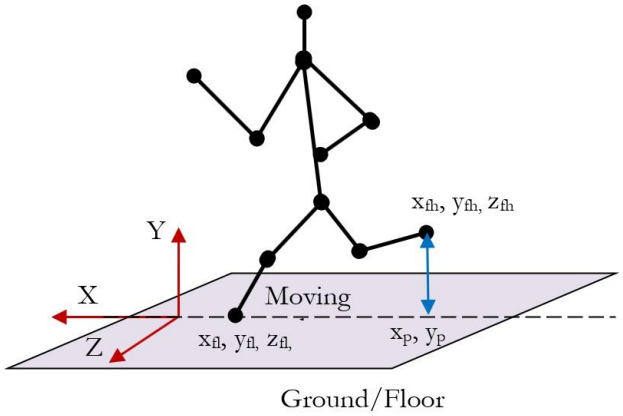


Fig. 4. Moving path and foot-lift in a movie frame.

First, it needs to find a short motion path on the ground/floor. It is considered as a linear line since it is a short path during a small number of frames. Therefore, linear equation was used. To find the moving path equation, the coordinate (x_{f1}, y_{f1}, z_{f1}) was used. At least three coordinate points are necessary to develop a linear equation. Therefore, at least five consecutive frames were used to find moving path equation. Here, only 2D (XY or ZY, currently XY in Fig. 4) path line is considered because human motion will be in one direction during a very short time.

$$y_{fl} = \alpha_1 x_{fl} + \alpha_2 \tag{1}$$

where, x_{fl}=[x_{fl1}, x_{fl2}, x_{fl3},…, x_{fln}], y_{fl}= [y_{fl1}, y_{fl2}, y_{fl3},…, y_{fln}] are vectors of coordinates of lower foot from n consecutive video frames. In current study, the number of frames, n, is set as 5, 7, 9, 15 and 20. 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.

$$y_p = \alpha_1 x_p + \alpha_2 \tag{2}$$

$$\Delta y_h = y_{fh} - y_p \tag{3}$$

$$\Delta y_l = y_{fl} - y_p \tag{4}$$

$$R_{fl} = \frac{\Delta y_l}{H_m} \tag{5}$$

$$R_{fh} = \frac{\Delta y_h}{H_m} \tag{6}$$

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 lower 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, 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, temporal foot-lift features were calculated by means of the following equations.

$$R_{f,max} = \max([R_{fh}]) \tag{7}$$

$$E_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R_{fhi} - R_{fli}\right)^2} \tag{8}$$

where, $R_{f,max}$ is maximum normalized foot-lift, E_{rms} is root-mean-square error between normalized lower foot-lifts and normalized higher foot-lifts.

In this work, the video recording frame rate is 30 fps. Therefore, the time duration is 0.167~0.667 s for 5~20 frames respectively. For this work, pixel motion features were computed as follows:

$$PMI_y = \frac{1}{H_m}\sum_{i=1}^{n}\left|y_{fl}(i) - y_{fl}(i+1)\right| + \left|y_{fh}(i) - y_{fh}(i+1)\right| \tag{9}$$

$$PMI_{x,y,z} = \frac{1}{H_m}\sum_{i=1}^{n}\left|x_{fl}, y_{fl}, z_{fl}(i) - x_{fl}, y_{fl}, z_{fl}(i+1)\right| + \frac{1}{H_m}\sum_{i=1}^{n}\left|x_{fh}, y_{fh}, z_{fh}(i) - x_{fh}, y_{fh}, z_{fh}(i+1)\right| \tag{10}$$

It is a new version of pixel motion feature (PMI). Originally, the PMI is 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 were used. The first one is the cumulative sum of absolute vertical displacements of feet during n frames (5, 7, 9, 20 frames). The second one is the cumulative sum of absolute 3D displacements of feet during n frames (5~20 frames or 0.167~0.667 s). The features are normalized using H_m for fairness.

3.5. Classifiers

In this study, five different classifiers; Decision Tree classifier with maximum 20 splits, weighted KNN classifier with k=10, SVM classifier with a quadratic kernel function, Naïve Bayes classifier and Random Forest classifier were used. Since these classifiers are well-known in previous works [32, 33], the detailed theoretical discussion for each classifier is skipped out to find the compactness.

4. Experimental Results and Discussion

Using proposed method, the experimental tests were conducted for "Walking" and "Running" actions in four different datasets. First, 50 % of the observations was trained using five different classifiers. Then, the other 50% of observations was used for testing the performance of proposed method. The number of trained and tested data are shown in Table 1. The observations in each dataset were tested in order to observe the compatibility of the model to each dataset. The calculations were performed by using MATLAB environment. The computing device has a RAM 8.00 G memory and 2.4 GHz processing speed.

Table 1. Number of trained and test data.

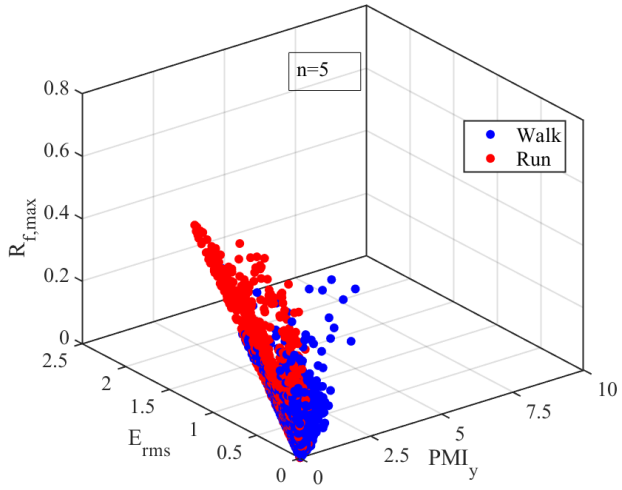
n	Trained/ Test Data	Action	UTKinet Dataset	KARD Dataset	G3D Dataset	CMU MoCap Dataset	Total
3 Frames	Trained data	**Run**	-	-	297	843	1140
		Walk	139	1060	315	835	2349
	Test data	**Run**	-	-	299	844	1143
		Walk	140	1066	317	833	2356
5 Frames	Trained data	**Run**	-	-	178	503	681
		Walk	85	634	184	496	1399
	Test data	**Run**	-	-	177	504	681
		Walk	83	637	187	498	1405
7 Frames	Trained data	**Run**	-	-	121	358	479
		Walk	59	450	136	358	1003
	Test data	**Run**	-	-	123	357	480
		Walk	57	452	137	356	1002
9 Frames	Trained data	**Run**	-	-	98	275	373
		Walk	43	352	100	275	770
	Test data	**Run**	-	-	97	277	374
		Walk	44	350	101	277	772
15 Frames	Trained data	**Run**	-	-	53	161	214
		Walk	22	207	54	163	446
	Test data	Run			55	163	218
		Walk	26	208	56	164	454
20 Frames	Trained data	Run			40	124	164
		Walk	20	151	44	121	336
	Test data	Run			41	122	163
		Walk	19	153	43	122	337

DOI:10.4186/ej.2024.28.7.41

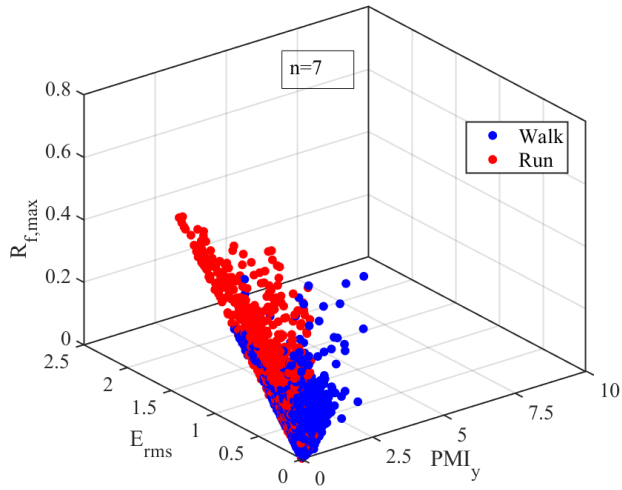
4.1. Feature Distribution

Indeed, visualizing the feature distribution is important for the selection of appropriate classifiers. It also helps in understanding the effect of the number of frames. Figure 5 (a)-(e) shows the 3D distribution of three temporal features, PMI_y, E_{rms}, $R_{f,max}$ using different number of frames. It should be noted that when the number of frames used in each recognition phase is smaller, the number of observations in each video clip is larger. Therefore, the number of observations decreases with the increment of the number of frames. At the same time, the separation of feature data becomes clear when the number of frames is larger. However, there are still mixing parts of features. It can be explained as follows. In running action, there are some moments when both feet are near to the ground and the foot-lift features in these conditions are similar to the features of walking action. In addition, some persons performed running action as jogging and the foot-lifts are not obvious and not much different with the foot-lifts of walking action. There is another reason that when training the recorded action video clips, there is no obvious foot-lifts in starting frames (start of action) and ending frames (end of action).

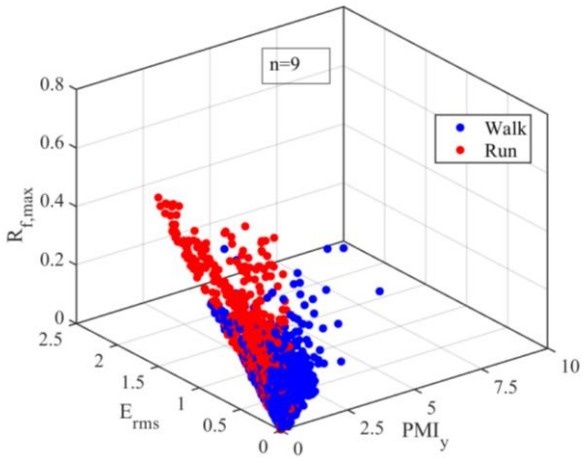
On the other hand, even in walking action, there are some persons whose foot-lifts are higher than normal due to their habit. For those conditions, all foot-lift features are relatively large. These are some reasons for why some parts of features still mixing.



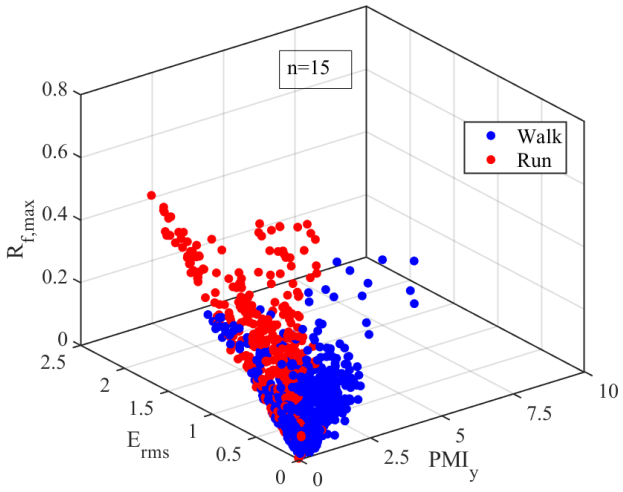
(a)



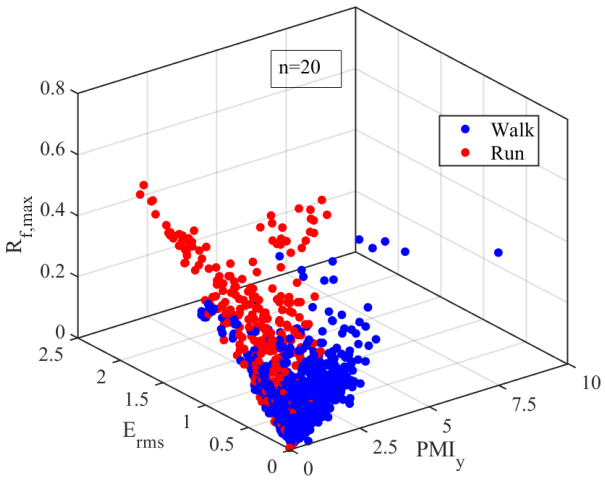
(b)



(c)



(d)



(e)

Fig. 5. Distribution of features; PMI_y, E_{rms} and $R_{f,max}$ (a) 5 frames, (b) 7 frames, (c) 9 frames, (d) 15 frames, (e) 20 frames.

4.2. Recognition Performance

The recognition performances (5-fold cross-validation accuracies and test accuracies) in different scenarios are list in Table 2 altogether. From these experimental results, the performance related to the number of frames used, the performance related to classifiers, the performance related to dataset and the performance related to the motion action can be understood.

Here, 5-fold cross-validation accuracy is the accuracy obtained by randomly dividing the training data into 5 equal subsets and each subset is used to validate the classification model during training process. By looking at the cross-validation accuracy, the performance of trained classifier can be estimated for future tests.

First, it can be seen that the recognition performance in every test scenario increases when the number of frames used increases.

When looking at the performances of different classifiers in comparison, The KNN classifier gives the highest recognition accuracy compared to other classifiers at every test scenario. Then, there is a comparative recognition accuracy between SVM classifier and Random Forest classifier. The Decision Tree classifier is in the fourth place and Naïve Bayes classifier shows the lowest performance. The SVM classifier can reach to 100 % accuracy when using 20 frames. In KNN classifier 10 weighted nearest distances are considered for decision. It means that ten possible votes of action are taken into account in decision of classification. It is a good scheme for mixing classes. It makes KNN classifier best fit with proposed method. The other classifiers are suitable to use when the classes are distinctly separated.

Table 2. Performance of proposed method for different number of frames.

Classifier	5-Fold Cross Validation 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)
n=5 Frames							
Decision Tree	84.24	95.15	96.78	96.30	57.63	97.80	78.85
KNN	**86.16**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
SVM	84.94	83.64	94.43	95.44	50.85	99.20	66.53
Naive Bayes	78.17	90.30	97.25	92.76	48.59	90.06	44.99
Random Forest	84.70	89.10	93.64	95.71	53.11	97.79	74.68
n=7 Frames							
Decision Tree	88.38	96.46	98.23	97.72	74.39	98.31	92.58
KNN	**90.68**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
SVM	88.58	87.61	95.80	96.58	56.09	97.89	80.39
Naive Bayes	78.65	91.15	98.45	93.54	51.22	89.87	45.10
Random Forest	89.43	92.04	98.12	96.96	54.47	99.15	80.39
n=9 Frames							
Decision Tree	85.96	91.95	95.85	98.51	71.50	98.60	89.53
KNN	**88.58**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
SVM	87.93	85.06	96.14	94.53	58.03	97.29	76.00
Naive Bayes	79.23	90.80	97.85	91.54	61.14	89.87	44.40
Random Forest	87.36	90.80	95.85	95.52	64.77	97.11	81.23
n=15 Frames							
Decision Tree	86.13	96.15	98.80	98.28	90.00	98.78	92.33
KNN	**90.49**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
SVM	88.86	86.54	97.59	97.41	64.55	96.94	76.69
Naive Bayes	80.53	94.23	97.83	93.10	64.55	90.21	47.24
Random Forest	88.85	94.230	96.87	95.69	69.10	96.94	84.36
n=20 Frames							
Decision Tree	90.10	97.30	98.37	97.67	88.89	98.35	97.94
KNN	**93.97**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
SVM	93.67	**100.00**	99.02	97.65	74.07	97.12	90.53
Naive Bayes	82.81	94.60	98.37	92.94	66.67	90.12	56.79
Random Forest	92.86	**100.00**	98.37	98.823	71.60	98.77	90.53

As shown in Figs. 5 (a)-(e), the foot-lift features $PMI_{x,y,z}$, PMI_y, E_{rms} and $R_{f,max}$ of "Walking" and "Running" actions become different when the number of frames becomes lager. For this reason, generally higher recognition accuracy can achieved using a larger number of frames in each recognition phase. However, accuracy of up to 100% can be achieved by using only 5 frames and KNN classifier.

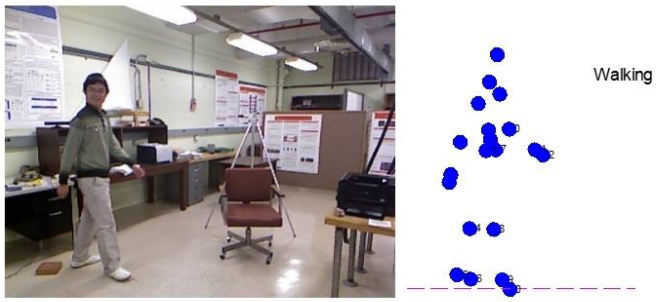
The results show that the performance of the proposed method is better for "Walking" action compared to "Running" action using Decision Tree, SVM, Naïve Bayes, and Random Forest classifier. However, when we use 20 frames in each recognition phase, the recognition accuracy for running becomes higher.

The reason behind it has been explained in previous section. In both trained and tested running action video clips, some frames are occupied by starting of action and

ending of actions in which there is no foot-lifts. For these conditions, the action is recognized as "Walking".

When comparing the recognition accuracy for each dataset, the proposed method relatively works well for "Walking" action of KARD dataset and "Running" action of MoCap dataset. Here, it should be noted that there is no effect of the number of joints available because only two foot joints information were used in this proposed method. In G3D dataset, "Running" and "Walking" actions are in-place action like motion on a treadmill. Since KNN classifier gives 100 % accuracy for all dataset, the proposed method is fit for all dataset. Some sample results are shown in Fig. 6.

In real-world applications, the proposed method is potential in human-following robots which require instantaneous recognition for "Walking" and "Running" actions. However, the proposed method requires a reliable method to exact the skeleton joint information for the feet.



(a)

(b)

Fig. 6. Visual observations for sample results (a) KARD dataset (b) G3D dataset.

4.3. Performance Comparison with Existing Methods

The performance of proposed method was compared with the performance of existing methods in terms of joint information requirement and frame requirement. The data are shown in Table 3. It should be noted that the datasets used in existing methods are different from each other. For most existing methods, the actions are other daily-live actions rather than "Walking" and "Running" actions. Also, most existing methods used at least 12 joints or the whole body joints. Meanwhile, the proposed method requires only two joints and 5 frames. Thus, the proposed method overcomes both challenges by reducing the joint information requirements and frame requirements.

Table 3. Comparison with existing methods.

Method	Joint information used	Number of frames used	Action	Accuracy (%)
[1]	12, 15, 39	125	Walking	100.00
[5]	The whole body	300	Other actions	95.00
[6]	The whole body	300	Other actions	92.08
[12]	The whole body	300	Other actions	96.30
[13]	The whole body	300	Other actions	96.30
[14]	The whole body	70-80	Other actions	99.81
Ours with KNN classifier	Two foot joints	5	Walking / Running	**100.00**

5. Conclusions

In this study, temporal foot-lift features were introduced and applied in recognition of "Walking" and "Running". This study is aimed at addressing spatial and temporal challenges; requirement of the whole-body joints information and the requirement of a large number of frames for motion action recognition. We used skeleton joint data from four popular datasets (KARD, UTKinet, G3D, CMUMoCap datasets) in training and testing processes. Five different classifiers including Decision Tree, KNN, SVM, Naïve Bayes, and Random Forest Classifier were used. The temporal foot-lift features were calculated by using only foot joint information in 5, 7, 9, 15 and 20 frames. The proposed method using temporal foot-lift features and KNN classifier can give accuracy of 100% for human motion action using 5 frames in each recognition phase. The performance of proposed method increases with the number of frames used in each recognition phase. When using other classifier rather than KNN, the recognition accuracy is higher for "Walking" action compared to "Running" action. The propose method is view-invariant, independent of starting location, starting frame as well as left/right specification. This work still have some limitations to be considered in future works. It should consider "Standing" and "Jogging" actions. Then it should also consider frame dropping strategy to reduce redundant information in processing.

Acknowledgement

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.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] 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, Dec. 2023,

[2] A. C. Cob-Parro, C. Losada-Gutiérrez, M. Marrón-Romera, A. Gardel-Vicente, and I. Bravo-Muñoz, "A new framework for deep learning video based Human Action Recognition on the edge," *Expert Syst Appl*, vol. 238, Mar. 2024.

[3] V. Jain, G. Gupta, M. Gupta, D. K. Sharma, and U. Ghosh, "Ambient intelligence-based multimodal human action recognition for autonomous systems," *ISA Trans*, vol. 132, pp. 94–108, Jan. 2023.

[4] K. Host and M. Ivašić-Kos, "An overview of Human Action Recognition in sports based on Computer Vision," *Heliyon*, vol. 8, no. 6. Elsevier Ltd, Jun. 01, 2022.

[5] 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.

[6] 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.

[7] K. Liu, L. Gao, N. M. Khan, L. Qi, and L. Guan, "Integrating vertex and edge features with Graph Convolutional Networks for skeleton-based action recognition," *Neurocomputing*, vol. 466, pp. 190–201, Nov. 2021.

[8] C. Dai, S. Lu, C. Liu, and B. Guo, "A light-weight skeleton human action recognition model with knowledge distillation for edge intelligent surveillance applications," *Appl Soft Comput*, vol. 151, pp.1–11, Jan. 2024,

[9] X. Li, Q. Huang, and Z. Wang, "Spatial and temporal information fusion for human action recognition via Center Boundary Balancing Multimodal Classifier," *J Vis Commun Image Represent*, vol. 90, pp. 1–13, Feb. 2023.

[10] H. Wang, B. Yu, K. Xia, J. Li, and X. Zuo, "Skeleton edge motion networks for human action recognition," *Neurocomputing*, vol. 423, pp. 1–12, Jan. 2021.

[11] Y. Tian, J. Chen, J. I. Kim, and J. Kwac, "Multiple-input streams attention (MISA) network for skeleton-based construction workers' action recognition using body-segment representation strategies," *Autom Constr*, vol. 156, pp. 1–16, Dec. 2023

[12] A. F. Bavil, H. Damirchi, and H. D. Taghirad, "Action Capsules: Human skeleton action recognition," *Computer Vision and Image Understanding*, vol. 233, pp. 1–11, Aug. 2023.

[13] 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.

[14] 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.

[15] 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

[16] 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.

[17] 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.

[18] 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.

[19] M. Feng and J. Meunier, "A Lightweight Graph Neural Network Algorithm for Action Recognition Based on Self-Distillation," *Algorithms*, vol. 16, no. 12, pp.1–12. Dec. 2023.

[20] G. Wang, P. Zhao, Y. Shi, C. Zhao, and S. Yang, "Generative Model-based Feature Knowledge Distillation for Action Recognition,", pp.1–9, Dec. 2023,

[21] Y. Chen, L. Wang, D. Hu, and H. Cheng, "Multi-view graph convolution network for the recognition of human action with spatial and temporal occlusion problems," *J Vis Commun Image Represent*, vol. 97, pp. 1–12, Dec. 2023.

[22] J. J. Seo, H. Il Kim, W. De Neve, and Y. M. Ro, "Effective and efficient human action recognition using dynamic frame skipping and trajectory

rejection," *Image Vis Comput*, vol. 58, pp. 76–85, Feb. 2017.

[23] N. ur R. Malik, U. U. Sheikh, S. A. R. Abu-Bakar, and A. Channa, "Multi-View Human Action Recognition Using Skeleton Based-Fine KNN with Extraneous Frame Scrapping Technique," *Sensors*, vol. 23, no. 5, pp. 1–23, Mar. 2023.

[24] N. Tasnim and J. H. Baek, "Deep Learning-Based Human Action Recognition with Key-Frames Sampling Using Ranking Methods," *Applied Sciences (Switzerland)*, vol. 12, no. 9, pp.1–18, May 2022,

[25] S. R. Mishra, T. K. Mishra, G. Sanyal, A. Sarkar, and S. C. Satapathy, "Real time human action recognition using triggered frame extraction and a typical CNN heuristic," *Pattern Recognit Lett*, vol. 135, pp. 329–336, Jul. 2020.

[26] H. Yasin, M. Hussain, and A. Weber, "Keys for action: An efficient keyframe-based approach for 3d action recognition using a deep neural network," *Sensors (Switzerland)*, vol. 20, no. 8, pp. 1–24. Apr. 2020.

[27] C. Yang, F. Mei, T. Zang, J. Tu, N. Jiang, and L. Liu, "Human Action Recognition Using Key-Frame Attention-Based LSTM Networks," *Electronics (Switzerland)*, vol. 12, no. 12, pp. 1–20. Jun. 2023

[28] 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.

[29] 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.

[30] 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.

[31] 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/

[32] F. H. dos S. Silva *et al.*, "A novel feature extractor for human action recognition in visual question answering," *Pattern Recognit Lett*, vol. 147, pp. 41–47, Jul. 2021.

[33] 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.

Khin Cho Tun is an Associate Professor, from Singapore-Myanmar Vocational Training Institute which is operating under the DTVET, Ministry of Science and Technology. She earned her bachelor degree in Electronic Engineering from Yangon Technological University. Then, she received her Master degree in Communication and Computer Network Technology from Assumption University (Thailand). Currently, she is conducting Ph.D study at Electronic Engienering Department, Yangon Technological University, Myanmar. She has published research papers in International conferences and journals inlcuding Engineering Journal. She has a good experience in research, teaching and supervising students' project works. She has achieved "Best Paper Award", in IEEE-KST Conference (2015) in Thailand.

Hla Myo Tun is a Leading Pro-Rector of Research and Engineering Higher Education of Yangon Technological University (YTU). He specializes in professional training for Engineering Higher Education leaders, heads of departments and faculty members. He also directs Research and Development programmes and workshops and works as a certified Quality Assurance Evaluator of Myanmar Engineering Council (MEngC) since 2019. Before promoting the Pro-Rector (Research) of YTU, Dr. Hla Myo Tun worked as a head of the Department of Electronic Engineering of Mandalay Technological University (MTU) and YTU in Myanmar. He was managing two research groups on Antenna Engineering and Semiconductor Electronics Engineering Technology. After earning his doctorate in Electronic Engineering in 2008, he led a Control System Design Research Lab in MTU. He did his research work on Time Reversal Focusing in Underwater Communication under Signal Processing Research Group of IIT Delhi in India in 2013. He got an award of Outstanding Research Report from IIT Delhi. Dr. Hla Myo Tun studied Semiconductor Device Fabrication and Properties Measurement (completing country focus trainings in Japanese Universities) from 2014 to 2019.

Khin Kyu Kyu Win, who is currently a professor, received the first degree, B.E (Electronics) in 1996 from Rangon Institute of Technology, Myanmar. She earned her second degree, Meng (Electrical and Electronics) in 2004 from Nanyang Tecnological University, Singapore. Her Ph.D (Electronics) was achived in 2018 from Yangon Technological Unveristy, Myanmar. Her specialied field is embedded system and wireless communication. In 1997, she joined the Department of Electronic Engineering at Yangon Technological University. She is amembership of Fedration of Myanmar Engineering Society.