A Basic Study on Ballroom Dance Figure Classification with LSTM using Multi-modal Sensor

Hitoshi Matsuyama, Kei Hiroi, Katsuhiko Kaji,
Takuro Yonezawa and Nobuo Kawaguchi

Abstract The paper presents a ballroom dance figure classification method with LSTM using video and wearable sensors. Ballroom dance is a popular sport among people regardless of age or sex. However, learning ballroom dance is very difficult for less experienced dancers as it has many complex types of "dance figures", which is a completed set of footsteps. Therefore, we aim to develop a system to assist dance exercise which gives advice proper to each dance figure characteristic by recognizing dance figures correctly. While the major approach to recognize dance performance is to utilize video, we cannot simply adopt it for ballroom dance because the images of dancers overlap each other. To solve the problem, we propose a hybrid figure recognition method combining video and wearable sensors to enhance its accuracy and robustness. We collect video and wearable sensor data of seven dancers including acceleration, angular velocity, and body parts location change by pose estimation. After that, we pre-process them and put them into an LSTM-based deep learning network. As a result, we confirmed that our approach achieved an F1-score of 0.86 for 13 figure types recognition using the multi-modal sensors with trial-based 5-fold cross-validation. We also performed user-based cross-validation, and sliding window algorithms. In addition, we compared the results with our previous method using Random Forest and also evaluated the robustness with occlusions. We found the LSTM-based method worked better than Random Forest with keypoint data. On the other hand, LSTM could not perform well with a sliding window algorithm. We consider the LSTM-based method would work better with a larger dance figure data, which is our next work. In addition, we will investigate how to solve occlusion problems with pose estimation.

Hitoshi Matsuyama, Kei Hiroi, Takuro Yonezawa, Nobuo Kawaguchi
Graduate School of Engineering, Nagoya University, Japan, e-mail: hitoshi@ucl.nuee.nagoya-u.ac.jp

Kei Hiroi
Disaster Prevention Research Institute, Kyoto University, Japan

Katsuhiko Kaji
Faculty of Information Science, Aichi Institute of Technology

1 Introduction

Ballroom dance is a popular sport among people regardless of age or sex. The dance is popular not only as a communication means, but also as a competitive sport to compete for the beauty, coolness, and perfection of performance. Through the dance, it is reported physical and cognitive decline are prevented [1]. While ballroom dance is enjoyed by many people, it is sometimes difficult for less experienced dancers as it has many complex types of dance figures.

Dance figure is a small sequence of footsteps comprising a meaningful gestalt, and every ballroom dance performance comprises a combination of dance figures. Among them, there are ones called "basic figure". Basic figure is the most cardinal movement of dance and given a specific name. Everybody who begins ballroom dance starts with learning how to step basic figures, then work hard to improve movement, posture, and musicality in each figure.

The most popular way to learn ballroom dance is to take a lesson at a dance studio. The main forms of a dance lesson including ballroom dance are (1) lessons taught by instructors, and (2) individual exercise by looking at the players themselves. In each practice form, several methods have been proposed to improve dance skills by assisting the dance exercise, such as transmitting the movement of the instructor to the recipient [2, 3, 4], or the system itself that plays the role of instructor [5, 6, 7]. However, those systems are not handling the information of dance figures.

In dance, especially ballroom dance, understanding figure types is an important method to clarify how to improve the performance because each figure type has its correct way of dancing. Ballroom dance has many complex figures in which the directions, timings of the footsteps, and orientation of the body are defined. Dancers learn and practice following those guidelines. For less experienced dancers, however, it is very difficult to remember and understand the figure types and guidelines. Therefore, we aim to assist basic figure learning and understanding with the automatic figure recognition and giving advice following each figure guideline. Automatic dance figure recognition may also lead to some other applications such as supporting dance coachers in a group lesson or creating a dance video to lecture basic figures.

In our previous works [8, 9], we have shown the possibility of ballroom dance figure classification using video and wearable sensors by extracting some basic feature values to put into Random Forest. However, as ballroom dance figure is a kind of time-sequential data, it is important to recognize characteristics of time sequence in each figure. Therefore, we adopt an LSTM, which is one of the major approaches to handle sequential data of human activities in these days, to develop an LSTM-based approach to classify the ballroom dance basic figures.

In this paper, we first collect ballroom dance figure data. We use wearable sensors attached to six body parts of a dancer to collect acceleration and angular velocity data. In addition, video data including two different shooting directions are acquired. We collect over 2 hours of data including 13 figure types, 2660 figure performances in total, from 7 different dancers. From the video data, we extract body keypoint data using OpenPose [10]. After that, we preprocess each data and put the sequential data in LSTM for each segmented single figure data or overlapped data with the sliding

window algorithm. The accuracy of figure classification is evaluated mainly by the F1 score. In addition, we compare the method with our previous works [8, 9] and also evaluate the robustness of our LSTM-based method and Random Forest-based method for occlusion. This work extends our previous works [8, 9] to an LSTM-based method, shows base-line results compared with Random Forest-based method, and evaluated the robustness of both LSTM-based and Random Forest-based method with occlusions.

As a result, we confirmed that our approach achieved an F1-score of 0.86 for 13 figure types recognition using the multi-modal sensors with trial-based 5-fold cross-validation. We also performed user-based cross-validation, and sliding window algorithms. In addition, we compared the results with our previous method using Random Forest and also evaluated the robustness with occlusions. We found the LSTM-based method worked better than Random Forest with keypoint data. On the other hand, LSTM could not perform well with a sliding window algorithm. We consider the LSTM-based method would work better with a larger dance figure data, therefore one of our next work is to enlarge the dataset.

2 Related Works

In this section, we will show some related works of dance performance recognition and supporting methods. A major approach is to transmit the posture or movement information of an instructor to a participant. For example, Fujimoto et al. proposed a visual-based system to support dance exercise [2] using Kinect. In the system, participants can know how to move their bodies by looking at the skeleton location of the instructor, which is overlapped onto the participants' images. In addition to using Kinect, Yamauchi et al. utilized a wireless-mouse, and developed a more accommodating dance supporting system [3]. There are also some footwear-based supporting approaches. Narazani et al developed a dance-skill transfer system by foot-base interaction [4]. Other footwear devices developments are known too [11, 12].

On the other hand, some works aim to construct a system that plays the role of instructor. For example, Anderson et al. developed an augmented mirror to support ballet exercise [5]. Milka et al. took their focuses on an augmented mirror too, and developed a visual and verbal feedback system for augmented mirror [6]. Not only designing an augmented mirror, but there is also a work that constructed a virtual instructor. Huang et al. [7] analyzed the ballroom dance lesson system, and divided the lesson time into some parts to develop a virtual ballroom dance instructor system.

Not only dance, but there are also many works, of course, to assist sports with a computer system, such as soccer [13], rugby [14] , and swimming [15]. The computational system assisting sports is reported to have a good effect on the participants [16].

What can be said in common to these works is they aim to improve general dance performance skills. However, as ballroom dance has many sorts of basic dance figures each of which has its respective guideline, it is important to give advice which is

Table 1 Dance figure names and their characteristics

	Number of foot actions	Progressing direction	Change amount of body orientation	Amount of free arm
OpenBasic	3 times	F	None	Free
FootChange	3 times	F and B	None	Free
Fan	3 times	S	90 degree ACW	Free
HockyStick	3 times	F and B	90 degree CW	Free
NewYorkR	3 times	F	90 degree CW	Free
NewYorkL	3 times	F	90 degree ACW	Free
SpotTurn	3 times	F	360 degree ACW	Free
NaturalTop	3 times	S	315 degree CW	Holding
OpeningOut	3 times	S	None	Holding
Alemana	3 times	F and B	90 degree CW	Free
HandtoHandR	3 times	B	90 degree ACW	Free
HandtoHandL	3 times	B	90 degree CW	Free
Aida	2 times	B	None	Free

F: Forward, B: Backward, S: Side, CW: Clockwise, ACW: Anti clockwise

more specialized to each dance figure type. Thus, in this paper, we develop a dance figure classification method for a ballroom dance learning assist system utilizing automatic dance figure recognition.

3 Ballroom Dance Dataset

In this section, we describe the ballroom dance figure dataset. As there was no ballroom dance dataset available which contains an inertial sensor and video data of several dance figures, we started the work with collecting the data. We asked seven experienced ballroom dancers to perform 13 types of dance figures. The names and other characteristics of each figure are stated in Table 1 and Fig. 1 shows how each figure is performed to music. As shown in Fig. 1, a dancer moves and steps with respect to each beat value. e.g., In Open basic a preparation movement starts from count 3, the right foot steps forward at count 4, whose movement continues until count 1, and the left foot steps forward at count 2.

In the dataset, a series of dance performances (routine) comprised of those 13 types of dance figures are given, and our 7 participants perform the routine. The order of figure types is “OpenBasic, FootChange, Fan, HockyStick, NewYorkR, NewYorkL, NewYorkR, SpotTurn, OpenBasic, NaturalTop, OpeningOut, FootChange, Fan, Alemana, HandtoHandR, HandtoHandL, HandtoHandR, Aida, SpotTurn". The attributes of the 7 dancers are shown in Tab. 2. The experiences of all of them are not less than 1 year, which means that all of them can perform the dance steps almost correctly. The height and experience of dancers vary from 160cm to 182cm, and 1 year to seventeen years. You can see that we collected only male performances. This is because the male’s and female’s steps are usually totally different in the same figure name.

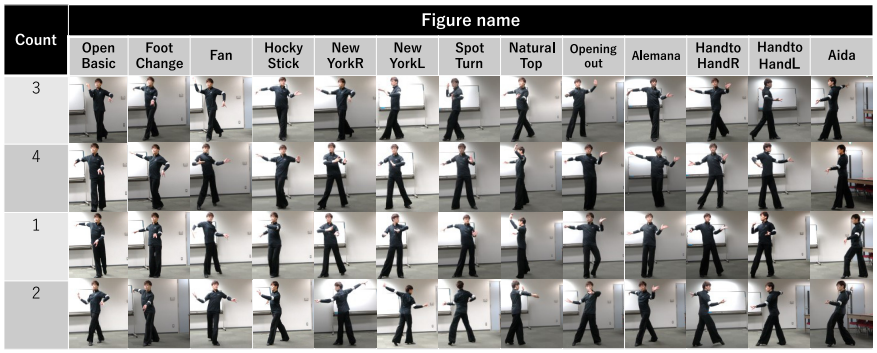

Fig. 1 Dance figures and movements with respect to music counts

Table 2 Property of dancers participated

	Sex	Height (cm)	Experience (year)
Dancer 1	Male	173	17
Dancer 2	Male	176	5
Dancer 3	Male	182	1
Dancer 4	Male	160	1
Dancer 5	Male	171	4
Dancer 6	Male	175	3
Dancer 7	Male	177	5

Table 3 Video and sensors for recording, positions, and the number of trials

	Number of data	Location	Sampling rate	Video/Sensors
Video	20 times	Arms, hips, and ankles	120 fps	Full HD(1920 x 1080)
Wearable sensor	20 times	Two positions	120 Hz	Accelerometer, gyroscope

While performing, video and wearable sensor data are acquired. Fig. 2 and Tab. 3 show how the data including two shooting directions are acquired. As shown in the figures, six wearable sensors (ATR-Promotions, TSND151, sampling rate = 120Hz) are worn on arms, hips, and ankles and the two shooting directions are from the center and from the back.

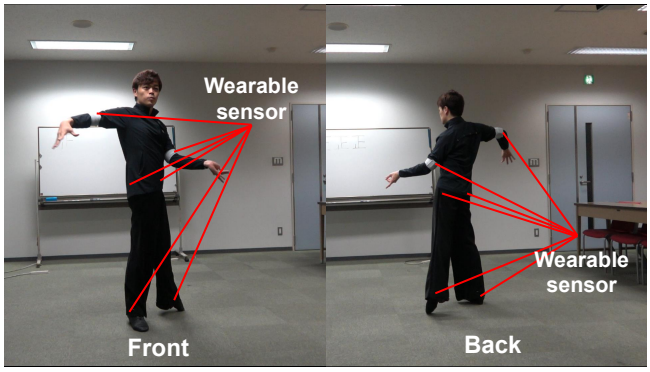

Fig. 2 Positions of worn sensors and shooting directions

4 Classification method using LSTM

The workflow of this study is shown in Fig.3. In order to perform ballroom dance figure classification using LSTM, we first preprocess each data. Then we reshape each figure data for LSTM input. After that, several forms of layers including the LSTM layer for deep learning classification model are designed. In final we put 6-dimension data from a wearable sensor and 50-dimension data from OpenPose into the classification model to get the result.

4.1 Preprocessing

The first step of classification is to preprocess data. The processes we adopt are as follows:

- Elimination of the movement distance in a timestep:
  The dancer changes position within the video frame during the performance. However, the movement distance and direction differ depending on the person, and also depending on the start position, thus we eliminate such effects. We first correct the coordinates of other body parts to relative positions based on the neck coordinate. Then we subtract each joint coordinate in each time from its beginning position in the timestep window.
- Interpolating missing value:
  Sometimes there are missing values in the data. To handle them, we utilize the interpolate function of the pandas in python.
- Standardization:
  In order to rescale the different types of modalities, we adopt standardization. Every data of every modal is rescaled to have a mean of 0 and a standard deviation of 1.
- selecting wearable sensor:

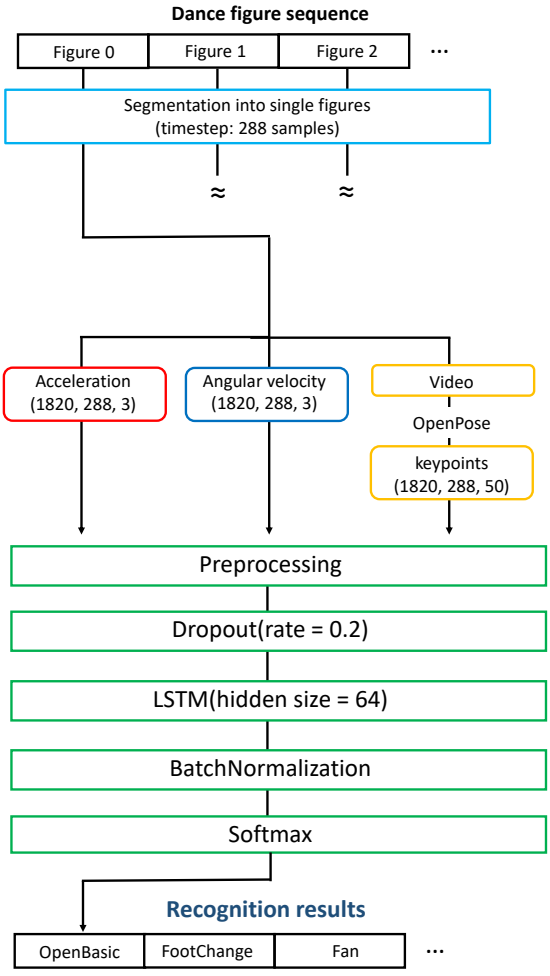

Fig. 3 Dance figure recognition method overview

Among six wearable sensors on right and left elbows, waists, and ankles, we tested which part help the classification [9]. Finally, we selected the left ankle.

4.2 Input data form

The preprocessed data are then segmented into single dance figures. The segmentation method is shown in Fig.4. As shown in the figure, we first calculate how many samples are in each dance figure from BPM(beat per minute) information and

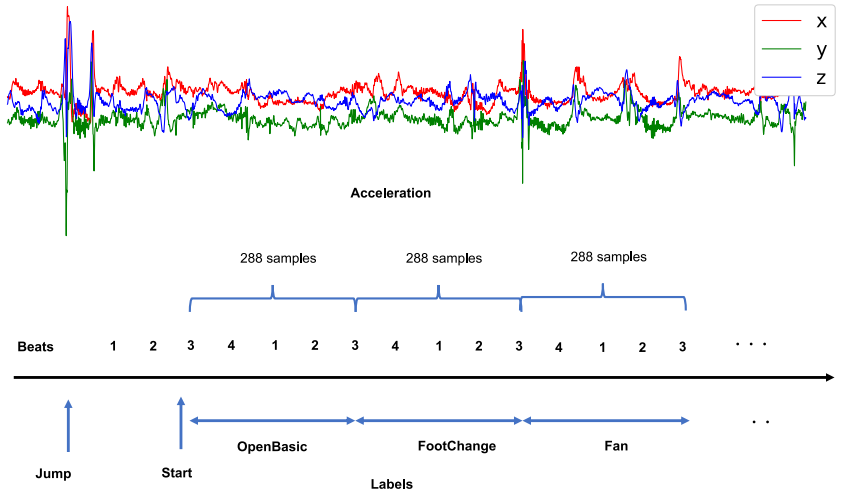

Fig. 4 Auto dance figure labeling method

the number of figures in the performance. In the Ballroom Dance Dataset, all of the performances are danced to the music of BPM 100. On the other hand, every sequence contains 19 figures(comprises of 13 figure types) in total, each of which has four beats. From this information and sampling rate, the number of samples in each dance figure is calculated as follows:
$samples = (60[sec.] \times 120[Hz]) \div (100[BPM] \div 4[beats\ per\ a\ figure]) = 288$
Where 60[sec.] $\times$ 120[Hz] calculates how many samples are there in 1 minute, 100[BPM] $\div$ 4[beats] calculates the number of sets of 4 beats(i.e. beats in a figure) in 1 minute. Thus, the formula provides the number of samples in each dance figure.

4.3 Designs of layers

After forming data, we design the deep learning network layers. As every ballroom dance figure has its characteristic sequence of footstep combination and body movement, LSTM-based classification is expected to provide good performance. Therefore we adopt LSTM and insert a dropout layer and batch normalization to prevent overfitting and internal covariate shift. An overview of the work process including DNN layer has shown in Fig.3, and Fig.5 shows the detail of how LSTM layer works after processed through Dropout layer. As shown in Fig.5, 6 dimension data from a wearable sensor(acc(x, y, z), gyr(x, y, z)) and 50 dimension data of keypoints from OpenPose [10, 17] of each segmented single figure are given as input data to LSTM.

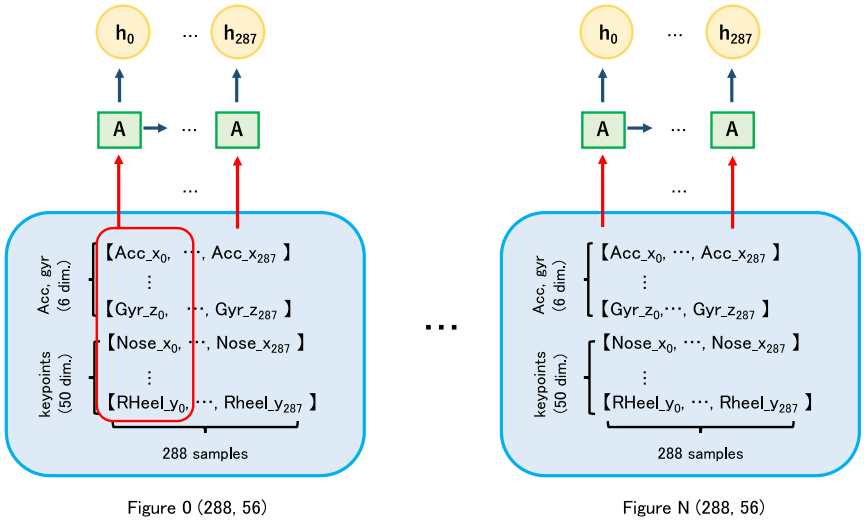

Fig. 5 LSTM layer structure and inputs for each timestep

5 Result

In this section, we first show a classification result using only wearable sensor data. Then we give a result of OpenPose keypoints data, and in final we show a result of the hybrid method utilizing multi-modal sensor data. For each modality of data, we first show a confusion matrix of 20% test data by a trial-based hold-out method. After that, we perform a trial-based and user-based 5-fold cross-validation method to evaluate its generalization ability. As some figure classes appear twice in the sequence, the numbers in each confusion matrix are divided by 2.

5.1 Acceleration and angular velocity

Tab.4 shows how our method worked with acceleration and angular velocity data for the 20% test data. Tab.7 shows the result of cross-validation methods. The Accuracy is 0.71 for trial-based 5FCV and 0.52 for user-based, while the F1-score is 0.68 for trial-based and 0.50 for user-based.

Table 4 Confusion matrix using acceleration and angular velocity data (trial base)

		Predicted												
		OpenBasic	FootChange	Fan	HockyStick	NewyorkR	NewyorkL	SpotTurn	NaturalTop	OpeningOut	Alemana	HandtoHandR	HandtoHandL	Aida
Answer	OpenBasic	25	0	1	0	3	0	0	0	0	0	0	0	0
	FootChange	0	18	0	2	0	0	0	0	0	2	1	0	0
	Fan	16	2	12	0	2	0	0	0	0	3	0	0	0
	HockyStick	0	2	0	10	0	0	0	1	1	7	0	6	0
	NewyorkR	2	1	2	0	19	0	1	2	0	0	0	0	0
	NewyorkL	0	0	0	0	0	20	6	1	0	0	0	0	0
	SpotTurn	0	0	0	0	0	0	21	0	0	0	0	0	0
	NaturalTop	0	0	0	0	0	0	0	32	0	0	0	0	2
	OpeningOut	1	0	0	0	1	0	1	4	18	0	0	0	1
	Alemana	2	2	2	12	1	0	0	0	0	9	0	2	0
	HandtoHandR	0	1	0	3	0	0	0	0	1	0	30	2	0
	HandtoHandL	0	1	0	0	0	0	0	1	0	2	0	19	2
	Aida	0	1	0	0	0	0	2	1	1	0	0	0	15

Table 5 Confusion matrix using keypoint data (trial base)

		Predicted												
		OpenBasic	FootChange	Fan	HockyStick	NewyorkR	NewyorkL	SpotTurn	NaturalTop	OpeningOut	Alemana	HandtoHandR	HandtoHandL	Aida
Answer	OpenBasic	23	0	1	0	0	0	0	0	0	1	0	0	0
	FootChange	0	20	0	0	0	0	0	2	2	0	1	0	0
	Fan	3	0	17	3	2	1	0	0	0	1	0	0	0
	HockyStick	0	2	0	13	0	0	0	1	1	7	0	2	0
	NewyorkR	2	1	2	0	15	2	1	2	0	0	6	2	0
	NewyorkL	0	0	0	0	3	17	6	1	0	0	1	3	0
	SpotTurn	0	0	0	0	0	0	22	3	0	0	0	0	0
	NaturalTop	0	0	0	0	0	0	1	23	0	0	0	1	1
	OpeningOut	1	0	0	1	1	0	0	2	21	0	0	0	1
	Alemana	1	1	0	7	1	0	0	1	0	14	0	1	1
	HandtoHandR	0	1	1	3	3	1	0	0	1	0	22	2	0
	HandtoHandL	0	1	0	1	1	2	0	1	1	2	0	20	2
	Aida	0	1	0	2	1	0	1	1	1	1	1	0	21

5.2 Keypoint from OpenPose

Tab.5 shows how our method worked with keypoint data from OpenPose for the 20% test data. Tab.7 shows the result of cross-validation methods. The Accuracy is 0.80 for trial-based 5FCV and 0.61 for user-based, while the F1-score is 0.80 for trial-based and 0.58 for user-based.

5.3 Hybrid method

Tab.6 shows how our method worked with hybrid method utilizing multi-modal sensor data for the 20% test data. Tab.7 shows the result of cross-validation methods. The Accuracy is 0.87 for trial-based 5FCV and 0.66 for user-based, while the F1-score is 0.86 for trial-based and 0.63 for user-based.

Table 6 Confusion matrix using acceleration, angular velocity, and keypoint data (trial base)

		Predicted												
		OpenBasic	FootChange	Fan	HockyStick	NewyorkR	NewyorkL	SpotTurn	NaturalTop	OpeningOut	Alemana	HandtoHandR	HandtoHandL	Aida
Answer	OpenBasic	26	0	1	1	1	0	0	0	0	0	0	0	0
	FootChange	0	20	0	1	0	0	0	0	0	0	1	0	0
	Fan	2	2	22	0	2	0	0	0	1	0	0	0	0
	HockyStick	1	2	2	12	0	0	0	1	0	5	0	2	0
	NewyorkR	0	1	1	0	17	1	1	0	0	0	0	0	0
	NewyorkL	0	0	0	0	0	20	6	1	0	0	0	0	0
	SpotTurn	0	0	0	0	0	0	21	0	0	0	0	0	0
	NaturalTop	0	0	0	0	0	1	0	25	2	0	0	0	2
	OpeningOut	1	0	0	0	1	0	1	0	21	0	0	0	1
	Alemana	0	0	0	5	1	0	0	0	0	20	0	1	0
	HandtoHandR	0	1	0	2	0	1	0	0	1	0	28	1	2
	HandtoHandL	0	0	0	0	0	0	0	1	0	0	0	24	0
	Aida	0	0	0	0	0	0	0	0	1	0	0	0	19

6 Evaluation

6.1 Overall evaluation

First, we evaluate our LSTM-based classification method with trial-based and user-based 5-fold cross-validations. Tab.7 summarizes the results of each experiment. Selecting Accuracy and F1-macro score as evaluation indicators, we calculated them for each modality while changing percentages of overlapping timesteps.

What is notable is that we achieved over 0.8 for both Accuracy and F1-score when we use keypoint data and hybrid method without overlapping timesteps. It means our LSTM-based method can recognize the segmented single figures with the auto figure labeling method for over-0.8 correctness.

6.2 User-based and trial-based cross-validation

As cross-validation methods we adopted trial-based and user-based metrics. Trial-based metric supposes situations where the recognition system can obtain data of a target user in advance for training(e.g. the system first asks the user to perform a specific sequence). On the other hand, user-based metric supposes the system recognizes a completely new user without any training.

From Tab.7, we can see there are approximately 0.2 point differences between trial-based and user-based methods. While the best score of the trial-based method is 0.87, user-based achieved 0.66, which means our LSTM-based method finds it difficult to recognize figures of a completely new dancer.

6.3 Scores with the sliding window algorithm

In our previous work [9], we focused on recognizing segmented single figures. Segmenting dance a figure needs BPM and figure types beforehand. In real cases,

Table 7 Classification results with LSTM-based method

Modality			Acc, Gyr			Keypoint			Hybrid		
Overlap(%)			0	50	75	0	50	75	0	50	75
Trial base (5FCV)	Accuracy	Mean	0.71	0.69	0.43	0.80	0.70	0.54	0.87	0.82	0.72
		Std	0.024	0.024	0.020	0.025	0.022	0.048	0.028	0.026	0.040
	F1 macro	Mean	0.68	0.67	0.36	0.80	0.69	0.50	0.86	0.82	0.71
		Std	0.033	0.025	0.027	0.027	0.025	0.045	0.028	0.026	0.042
User base (5FCV)	Accuracy	Mean	0.52	0.51	0.34	0.61	0.56	0.46	0.66	0.63	0.53
		Std	0.089	0.093	0.037	0.086	0.101	0.077	0.113	0.116	0.066
	F1 macro	Mean	0.50	0.47	0.26	0.58	0.53	0.42	0.63	0.62	0.51
		Std	0.088	0.082	0.043	0.095	0.111	0.083	0.128	0.114	0.070

however, it is sometimes difficult to acquire those. For example, in a ballroom dance competition, every pair of dancers perform their own routines comprised of the unique combination of figures. In such cases, we need mini-batch or online dance figure recognition. Thus, we adopt a sliding window algorithm with 0, 50, and 75 percent overlaps while the timestep has always 288 samples.

Tab.7 shows the results of 0, 50, and 75 percent overlaps. In general, Accuracy and F1-score decrease when overlap extent gets wide. In particular, results of keypoint data get extremely low with 75 percent overlaps. On the other hand, the hybrid method showed strength with 50 percent overlaps, though the scores also decreased approximately 0.15 with 75 percent overlaps. The reason for the low accuracy of the 75 percent overlaps is that for some figures, there are multiple choices of figures that come before them and the model get confused. To handle this problem, we need to prepare more training data to tell the model the tendency of figures that comes before each figure.

6.4 Comparison with Random Forest-based method

Picking up the trial-based 5FCV results, we made a comparison with our Random Forest-based classification method with basic feature extraction such as mean and variance of each axis sequence of each modality data. This is an advanced version of our previous Random Forest-based work [9]. The results of Random Forest-based method and LSTM-based method are summarized in Tab.8. From the table, we can see the hybrid method using LSTM has performed much better than just using acceleration and angular velocity from a wearable sensor, or keypoints from OpenPose. Although the hybrid method using LSTM performed better than using single modal data, on the other hand, it did not work better than the Random Forest-based method. We suppose it is because the amount of data is small to apply LSTM, where each dance figure class has 140 times of performances, compared to some works on activity classification using deep learning method [18, 19].

Table 8 F1 score of Random Forest-based method v.s. LSTM-based method

Modality		Acceleration, angular velocity	Keypoints from OpenPose	Hybrid method
F1 score	RF-based	0.87	0.74	0.92
	LSTM-based	0.71	0.80	0.86

Table 9 F1 score of Random Forest-based method v.s. LSTM-based method with occlusion

Occlusion	Acc, Gyr		Keypoints		Hybrid	
	RF	LSRM	RF	LSTM	RF	LSTM
None			0.74	0.80	0.92	0.86
Right 1/2 body			0.63	0.65	0.88	0.78
Left 1/2 body	0.87	0.71	0.66	0.55	0.87	0.70
Right 3/4 body			0.62	0.45	0.87	0.71
Left 3/4 body			0.58	0.45	0.87	0.70

6.5 Random Forest v.s. LSTM with occlusion

Up until here, we handled a single dancer in each video data. However, in the real ballroom dance situation, we often get occlusions by their partner because the dance is usually performed by pairs of dancers. Therefore, we virtually generate occlusion by hiding body keypoints from OpenPose. Considering how ballroom dance is performed, we decided to hide right or left half or 3/4 of the body to reproduce vertical occlusion by the partner of a dancer.

Tab. 9 shows the result of the experiment. To compare with our previous method using Random Forest, we write together the results of our previous method and the LSTM method. From the table, we can see LSTM-based resulted in lower f1 scores than our previous method in every part. In particular, f1 score using acceleration and angular velocity is much lower, which caused low scores in the hybrid method too.

7 Discussion

7.1 Low accuracy problem with the sliding window algorithm

The main problem of LSTM-based method is the accuracy gets low when we apply sliding window algorithms with 75 percent overlaps. Although we have collected more than 2 hours of data in total, the number of dance figures in the dataset is about 2000, which is not enough to get as high accuracy as Random Forest-based method. The matter of data collection is the difficulty in developing and collecting different sequences of performance. In particular, if we want to collect new types of figures, we first “memorize” how to perform each figure and then how the figures are ordered. However, every ballroom dance club or studio has its own dance sequence,

or “routine”, therefore we can ask other dance groups to collect different types of figures. The third problem is that performing ballroom dance is much harder than general activities, which cannot be helped.

7.2 User-independent recognition

The results with user-based cross-validation were lower compared to the trial-based method. We found both trial-based and user-based has its use scenario. Trial-based metric supposes situations where the recognition system can obtain data of a target user in advance for training(e.g. the system first asks the user to perform a specific sequence). On the other hand, the user-based metric supposes the system recognizes a completely new user without any training. This is a more general case and easy to try for users. Our method should be improved to perform better in this case.

The main problem of user-independent recognition is the ways of performing each dance figure differs among dancers. In particular, foot-action characteristics diverse severely, which is the main reason for the low accuracy of user-based classification with wearable sensor data. Although directions and timings of steps for each figure are described in the textbook, usually dancers do not follow the guides perfectly. Moreover, timings of preparation movement before each step and foot speeds are different among dancers, which confuses LSTM classifier. Considering those problems, we need to apply some additional preprocessing or algorithms to exactly extract features common among dancers for each dance figure.

7.3 Occlusion problem

Handling with occlusion problems also needs more improvement. In this paper we utilized the strength of a wearable sensor and dealt with occlusions. However, we could investigate how to utilize the keypoint data. In particular, though the LSTM-based method maintained its accuracy over 0.5 or 0.6 without half body information, the hybrid method could not improve its accuracy with keypoints. This is because we simply combine the wearable sensor data and keypoint data as input and let the LSTM learn the importance of each feature. However, it is possible to obtain the reliabilities of each keypoint coordinate with pose estimation. Therefore, controlling how the LSTM weighs each keypoint data with its reliability by adding that information as input or adding an extra gate in the LSTM cell to control weights of each keypoint may help the classification with occlusions.

8 Conclusion

In this paper, we proposed a ballroom dance figure classification method with LSTM using video and wearable sensors. We adopted both video and wearable sensors to enhance the accuracy and robustness of dance figure classification. We collected video and wearable sensor data of seven dancers including acceleration, angular velocity, and body parts location change by pose estimation. After that, we pre-processed them and put them into the LSTM-based deep learning network. As a result, we confirmed that our approach achieved f1-score 0.86 for 13 figure types recognition using multimodal sensors. Also, we hide the parts of dancers to reproduce occlusions. We found the LSTM-based method worked better than Random Forest with keypoint data. On the other hand, LSTM could not perform well with a sliding window algorithm. From all of the results, we concluded that we need to enlarge the size of our dance figure dataset for better LSTM performance.

Acknowledgements This research is partially supported by JSPS Grant-in-Aid for Scientific Research (B) Grant Number 17H01762 and JST CREST Grant Number 18071264. The ballroom dance performances were provided by the members of Nagoya University Ballroom Dance Club and its alumni.

References

1. Dafna Merom, Robert Cumming, Erin Mathieu, Kaarin J. Anstey, Chris Rissel, Judy M. Simpson, Rachael L. Morton, Ester Cerin, Catherine Sherrington, and Stephen R. Lord. Can Social Dancing Prevent Falls in Older Adults? a Protocol of the Dance, Aging, Cognition, Economics (DAnCE) Fall Prevention Randomised Controlled Trial. *BMC Public Health*, 13(1):477, 2013.
2. Minoru Fujimoto, Masahiko Tsukamoto, and Tsutomu Terada. A Dance Training System that Maps Self-Images onto an Instruction Video.
3. Masashi Yamauchi, Ryo Shinomoto, Eriko Nishiwaki, Risa Onozawa, and Tetsuro Kitahara. Development of Dance Training Support System Using Kinect and Wireless Mouse. *The Sysmposium of Entertainment Computing*, 2013:332–338, 2013.
4. Marla Narazani, Katie Seaborn, Atsushi Hiyama, and Masahiko Inami. StepSync: Wearable skill transfer system for real-time foot-based interaction, 2018.
5. Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. YouMove: Enhancing Movement Training with an Augmented Reality Mirror. *In Proc. of UIST 2013 Conference:ACM Symposium on User Interface Software and Technology*, pages 311–320, 2013.
6. Milka Trajkova and Francesco Cafaro. Takes Tutu to Ballet: Designing Visual and Verbal Feedback for Augmented Mirrors. *In Proc. of ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):1–30, 2018.
7. Hung-Hsuan Huang, Masaki Uejo, Yuki Seki, Joo-Ho Lee, and Kyoji Kawagoe. Construction of a Virtual Ballroom Dance Instructor. *The Japanese Society for Artificial Intelligence*, 28(2):187–196, 2013.
8. Hitoshi Matsuyama, Kei Hiroi, Katsuhiko Kaji, Takuro Yonezawa, and Nobuo Kawaguchi. Hybrid Activity Recognition for Ballroom Dance Exercise using Video and Wearable Sensor. In *International Conference on Activity and Behavior Computing*, 2019.

9. Hitoshi Matsuyama, Kei Hiroi, Katsuhiko Kaji, Takuro Yonezawa, and Nobuo Kawaguchi. Ballroom Dance Step Type Recognition by Random Forest Using Video and Wearable Sensor. In *International Workshop on Human Activity Sensing Corpus and Application*, 2019.
10. Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-person 2D Pose Estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.
11. Paradiso Joseph, Hu Eric, and Hsiao Kai yuh. The CyberShoe: A Wireless Multisensor Interface for a Dancers Feet. 03 1999.
12. J. A. Paradiso, K. Hsiao, A. Y. Benbasat, and Z. Teegarden. Design and Implementation of Expressive Footwear. *IBM Systems Journal*, 39(3.4):511–529, 2000.
13. Reza Maanijou and Seyed Abolghasem Mirroshandel. Introducing an expert system for prediction of soccer player ranking using ensemble learning. *Neural Computing and Applications*, 31(12):9157–9174, Dec 2019.
14. Nikolai B. Nordsborg, Hugo G. Espinosa, and David V. Thiel. Estimating energy expenditure during front crawl swimming using accelerometers. *Procedia Engineering*, 72:132 – 137, 2014. The Engineering of Sport 10.
15. Mark Waldron, Craig Twist, Jamie Highton, Paul Worsfold, and Matthew Daniels. Movement and physiological match demands of elite rugby league using portable global positioning systems. *Journal of sports sciences*, 29:1223–30, 08 2011.
16. Hua-Tsung Chen, Yu-Zhen He, and Chun-Chieh Hsu. Computer-assisted yoga training system. *Multimedia Tools and Applications*, 77(18):23969–23991, Sep 2018.
17. Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
18. N. Dawar and N. Kehtarnavaz. Action detection and recognition in continuous action streams by deep learning-based sensing fusion. *IEEE Sensors Journal*, 18(23):9660–9668, Dec 2018.
19. I. Hwang, G. Cha, and S. Oh. Multi-modal human action recognition using deep neural networks fusing image and inertial sensor data. In *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 278–283, Nov 2017.