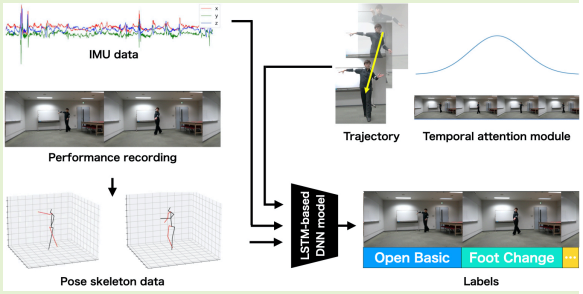
Deep Learning for Ballroom Dance Recognition: A Temporal and Trajectory-Aware Classification Model with Three-dimensional Pose Estimation and Wearable Sensing

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

***Abstract*— Dance performance recognition methods have been investigated and shown various applications such as picture-pose evaluation and synchronizing foot timing and direction. However, detailed analysis and feedback are still missing. To provide them, understanding the performance by component level is necessary. Specifically, we formulate it as a dance-figure classification problem using three-dimensional body joints and wearable sensors. Our model is based on long short-term memory (LSTM) and includes the temporal and trajectory-wise structure that uses the trajectory information in a timestep and the temporal masking module. As a result, we achieved 93% accuracy with our proposed method, which is highly overwhelming the baseline result (84.7%) and very close to the accuracy of the experienced dancers (93.6%). We have made the dataset of ballroom dance performance dataset open to researchers to develop the activity recognition field further.**



***Index Terms*— Activity recognition, Sensor systems and applications, Neural networks, Image motion analysis, Machine learning**

I. Introduction

BALLROOM dance is a popular sport among people regardless of age or sex. It is popular not only as a means of communication but also as a competitive sport for beauty, skill, and performance. Its effectiveness in preventing physical and cognitive decline by participating in the dance community has been reported [1]. However, though many people enjoy ballroom dancing, it is difficult to become a ballroom dance, especially for less experienced dancers, as it has many complex types of dance figures to learn and practice.

Dance figure is a small sequence of footsteps comprising a meaningful gestalt, and every ballroom dance performance contains a combination of dance figures. “Basic figure” is the most cardinal movement of dance, and each of which has a specific name. A beginner starts with learning the steps for basic figures and then works hard to improve movement, posture, and interpretation of music.

The most popular way to learn ballroom dance is to take a lesson at a dance studio. The primary forms of a dance lessons, including ballroom dance, are lessons taught by instructors and individual exercises by looking at the dancers themselves. In each practice form, researchers have proposed several methods to improve dance skills by assisting the dance exercise, such as transmitting the instructor’s movement to the recipient [2]–[4], or the system itself that plays the role of instructor [5]–[7]. However, these studies do not reach component-level recognition.

Understanding dance at the component-level is crucial to clarify how to improve the performance. In ballroom dance, dance figures are components, and each of which has its correct way of execution. 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 according to these guidelines. However, it is not easy for less experienced dancers to remember and understand the figure types and guidelines. Therefore, we developed an automatic recognition algorithm for dance figures to assist essential

Manuscript received April 24, 2021; accepted July 4, 2021. This work was supported in part by JSPS Grant-in-Aid for Scientific Research (B) Grant Number 17H01762, JST CREST Grant Number 18071264, and the DII Collaborative Graduate Program for Accelerating Innovation in Future Electronics at Nagoya University.

Hitoshi Matsuyama, Shunsuke Aoki, Takuro Yonezawa and Nobuo Kawaguchi are with Graduate School of Engineering, Nagoya University, Japan (e-mail: hitoshi@ucl.nuee.nagoya-u.ac.jp)

Kei Hiroi is with Disaster Prevention Research Institute, Kyoto University, Japan

Katsuhiko Kaji is with Faculty of Information Science, Aichi Institute of Technology, Japan

figure learning and understanding. Automatic dance figure recognition can help dance figure learning and allow dancers to analyze the trends of choreographies. In the future, it may also be possible to develop a support system for dance coachers in a group lessons or create dance videos to lecture basic figures.

In our previous works [8], [9], we showed the possibility of ballroom dance figure classification using video and wearable sensors by extracting some fundamental feature values to put into Random Forest. However, as the ballroom dance figure is time-sequential data, it is essential to recognize the time sequence characteristics of each dance figure. Therefore, our proposed methodology bases on long short-term memory(LSTM) [10], one of the effective approaches to handle sequential data of human activities, to develop an LSTM-based approach to classify the ballroom dance basic figures. In our latest work [11], we presented a study on ballroom dance figure classification with LSTM using two-dimensional(2D) pose data and wearable sensor data. Although we obtained 80%+ classification accuracy with a basic LSTM-based method, the method had room to improve the algorithm and structure.

In this study, we build a classification model to recognize the ballroom dance figures using a wearable sensor and a three-dimensional(3D) human pose estimation method [12]. We first preprocess the achieved time-sequential acceleration, angular velocity, and 3D pose data, followed by segmenting them into single dance figures. The segmented timesteps were processed through our LSTM-based deep neural network(DNN) to predict the dance-figure label. The label data for each segment were automatically generated with music rhythm, choreography information, and sampling rate. Nect, we obtained the trajectory information of the participant's middle hip and provided it to the classification model. Subsequently, we added wearable sensor data attached to the body of the participants. Each sensor position (right and left arms, waists, and ankles) was added to the 3D pose data one by one to choose the best position among the six available positions. In addition, we applied the temporal masking module to the dance figure sequence to subject higher importance to the characteristics of the figures.s

As a result, we achieved 93% accuracy with our proposed method, which is significantly better than the baseline result (84.7%) and very close to the accuracy of the experienced dancers (93.6%). We made the dataset of ballroom dance performance open to researchers to further develop the activity recognition field.

The main contributions of this paper are:

1) We built a dance-figure classification model using a 3D pose estimation method and make comparison with 2D joints.
2) We developed a hybrid structure of 3D joints and a wearable sensor.
3) We introduced a novel trajectory and temporal-aware structure.

In addition, we created a dataset of ballroom dance performance open to researchers to aid further development of the activity recognition field.

II. Related Work

In this section, we discuss works related to dance figure recognition. We first show the recent activity recognition methods, and then introduce the performance recognition and supporting methods.

A. Activity Recognition Method

With the growth of deep learning technology, several activity recognition researchers have introduced the technology to develop a new methodology [13]–[17]. Wang et al. [13] introduced an attention-based convolutional neural network for human activity recognition with wearable sensors. Mathews et al. [15] adopted a dictionary-learning approach. Vision-based deep learning approaches for activity recognition and prediction have also been investigated [18]–[20]. Villegas et al. [18] present a prediction model of the future video frame with an encoder-decoder convolutional neural network and convolutional LSTM.

However, the scope of these works is limited to the general activities such as walking, running, sitting; they are generally not modified to suit sports activities. Dance activities significantly differ from general activities as they involve active body movements, free arming, and rapid footsteps. Therefore, it is necessary to explore the features of dance and propose a novel activity recognition methodology.

B. Assisting Dance Performance and Other Sports

The dance performance recognition and supporting methods were also explored. A practical 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 exercises [2] using Kinect. Through the system, the participants can know how to move their bodies just by looking at the instructor's skeleton position that is overlapping the participants' images. In addition to using Kinect, Yamauchi et al. utilized a wireless mouse and developed a more accommodating dance support system [3]. Footwear-based device-based approaches were also explored [4], [21], [22]. Narazani et al. developed a dance-skill transfer system using a foot-base interaction [4].

Constrastingly, some works aimed to construct a system that played an instructor's role. For example, Anderson et al. developed an augmented mirror to support ballet exercises [5]. Milka et al. focused on an augmented mirror and created a visual and verbal feedback system for augmented mirrors [6]. In addition to designing an augmented mirror, a virtual instructor was also constructed. 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. Several researchers have investigated assisting sports with a computer system, such as soccer [23], rugby [24], and swimming [25]. The computational system helping sports is reported to positively affect the participants [26].

Although these works show promising results for their purpose, they aim to improve general dance performance skills. Ballroom dance has many types of basic dance figures,

TABLE I
BALLROOM 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

each of which has its respective guidelines. Therefore, it is important to give advice that is more specialized for each dance-figure type. Understanding these components also helps detailed analysis and serves entertainment purposes.

III. BALLROOM DANCE DATASET

In this section, we describe the ballroom dance-figure dataset. This dataset is collected by us that contains inertial sensor data and video data of dance figures. Seven experienced ballroom dancers participated in the data collection and performed a choreography that contained 13 types of dance figures. The names and other characteristics of each dance figure are listed in Table I. Figure 1 shows how we perform each figure to music beats. As shown in Figure 1, each dancer moves and steps to each beat count. For example, in Open Basic, a preparation movement starts from count 3, then the right foot steps forward at count 4, whose movement continues until count 1, and the left foot steps forward at count 2.

The choreography in the dataset comprises of the stated 13 figure types. The order of the figure types is "OpenBasic, FootChange, Fan, HockyStick, NewYorkR, NewYorkL, NewYorkR, SpotTurn, OpenBasic, NaturalTop, OpeningOut, FootChange, Fan, Alemana, HandtoHandR, HandtoHandL, HandtoHandR, Aida, and SpotTurn." Table II lists the heights and experiences of the participants. All of their experiences are not less than one year, and all of them can perform the dance steps almost correctly. The height and experience of dancers varied from 160cm to 182cm, and 1 to 17 years, respectively. We can see that the participants were only men. This is because the male's and female's steps are very different, even though they perform the same dance figure.

While performing, video and wearable sensor data were acquired. Figure 6 and Table III show how the performance data were acquired. As shown in the figures, six wearable sensors are worn on the right and left arms, hips, and ankles. The video directions were from the center and the back. We used TSND151[1] provided by ATR-Promotions. The sensor is small (40 × 50 × 14 mm) and light (27 g) enough to attach to a dancer and perform without any inconvenience. Each participant was equipped with sensors and performed the choreography. The sampling rate of the wearable sensor was adjusted to be the same as that of the video. We recorded 20 records of approximately 60 seconds of performance sequence using the sensors and video for each participant.

TABLE II
PROPERTY OF DANCERS IN THE BALLROOM DANCE DATASET.

	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

In the activity recognition field, it is sometimes difficult to collect specific kinds of data, such as dance activity. Thus, we created a ballroom dance dataset open to researchers in the activity recognition field for further development. The dataset is available on our website [2], and a description of the data are available on the Github page [3]. A sample video is also available [4]. The data available on the website include dance performance video, acceleration, angular velocity, and label data. We also prepared table data, including the participant, trial, figure label, and sensor modalities.

IV. METHODOLOGY

This section describes our method for ballroom dance figure classification using the 3D pose data of dancers. Figure 2 shows the workflow of the proposed method. First, we perform pose estimation and preprocessing. Then, we reshaped each figure data for the DNN structure. Figure 3 shows the overall network structure.

A. Pose Estimation

From the collected video of dance performance, we extracted 3D human pose data of a participant using VideoPose3D provided by FaceBook Research [12]. VideoPose3D first estimates the 2D pose data of a person and then performs the 3D reconstruction. Consequently, we achieved the 3D coordinates of the 17 joint positions of a participant.

B. Preprocessing

The processes we adopt are as follows:

- Elimination of the movement distance:
  The dancer changes their position within the video frame during the performance. However, the movement distance and direction differ depending on the person and depending on the start position. Thus, we eliminated such effects. Figure 4 shows how we eliminated the movement effect in a screen among frames in the same dance figure sequence, and how we added information of the

[1]https://www.atr-p.com/products/TSND121.html

[2]http://hub.hasc.jp/

[3]https://github.com/matsuyhit/BDD

[4]https://youtu.be/SV607W_ofGE

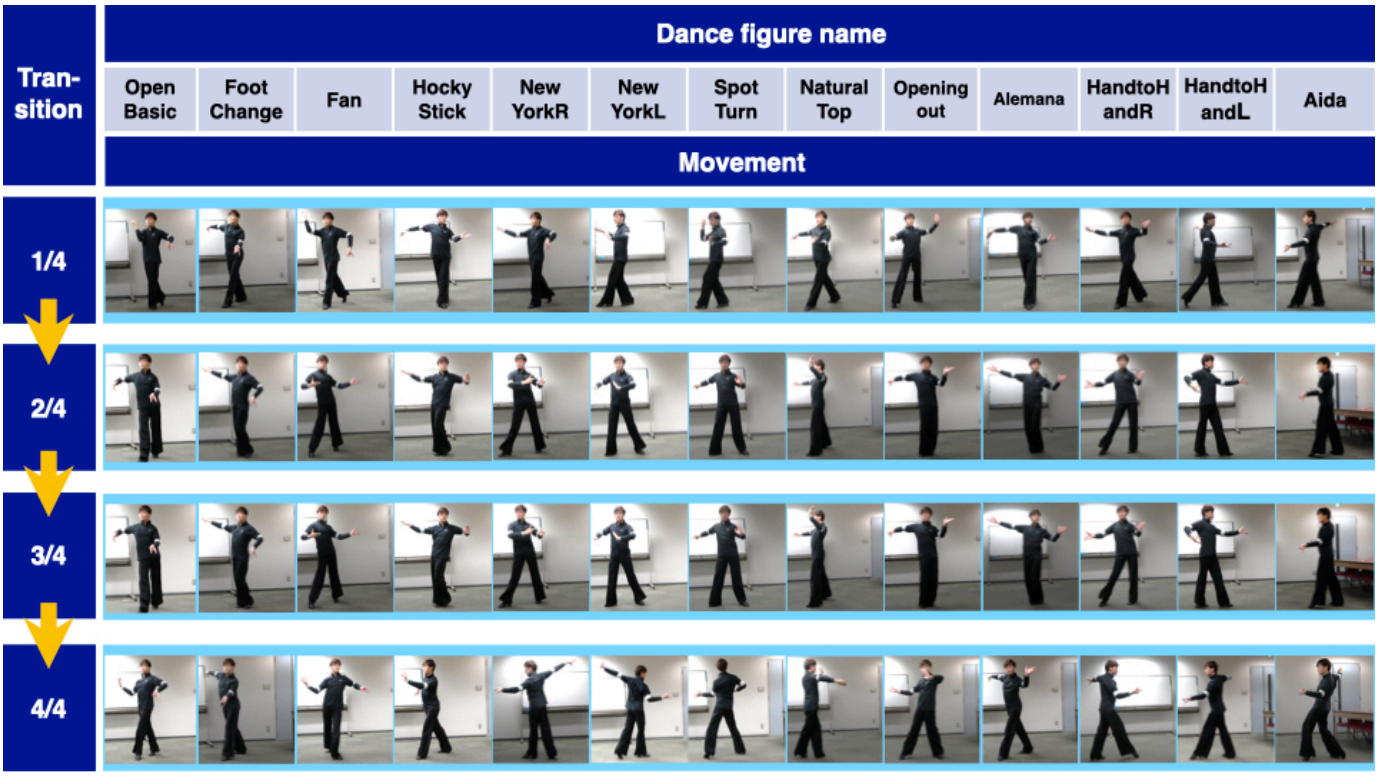


Fig. 1. Dance figures and movements for music counts.

TABLE III
VIDEO AND SENSORS FOR RECORDING, POSITIONS, AND THE NUMBER OF TRIALS.

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

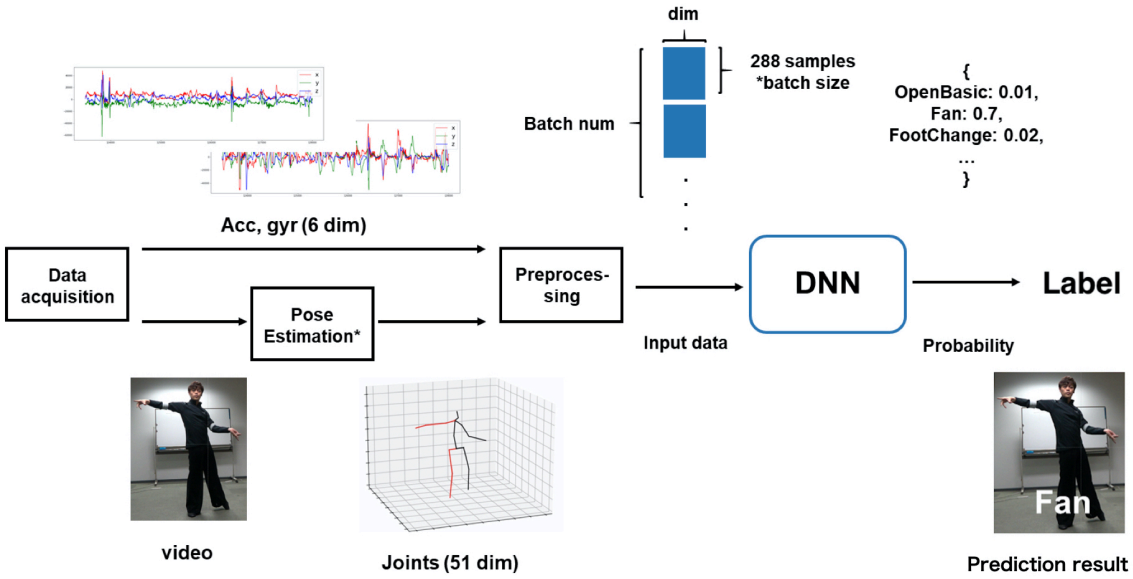


Fig. 2. Dance figure classification method.

movement vector. We set the middle-hip coordinate to the origin and changed the coordinates of the other body parts to the relative positions. Next, we added information on the pelvis coordinates' trajectory to give information

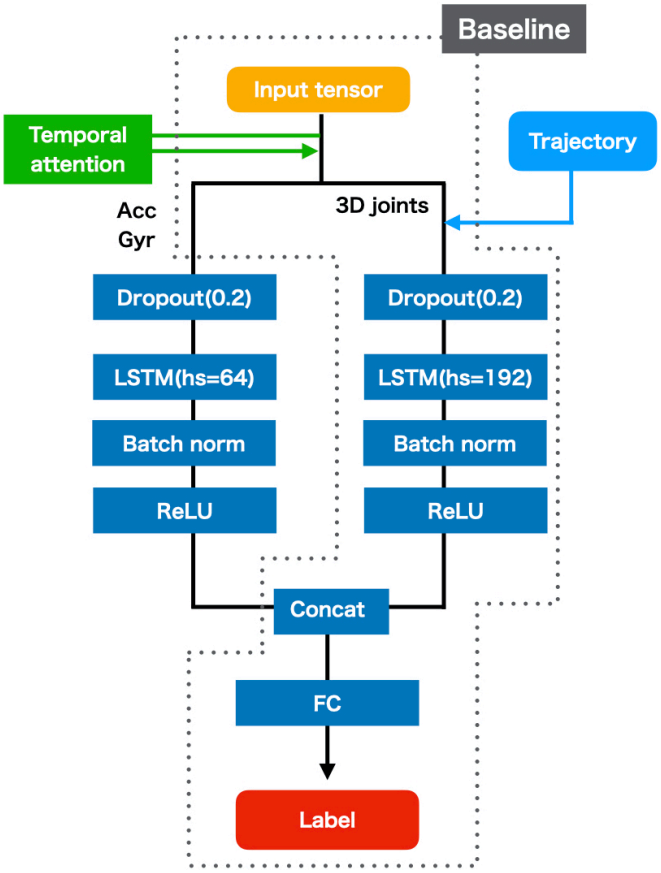


Fig. 3. Deep neural network structure for the baseline method and the hybrid of the 3D joints and wearable sensor. The gray dotted line shows the structure of the baseline method. Input tensor consists of sensor/3D joints dimension, timestep, and batch size.

on the movements during the technique for each time step.

- Standardization:
  To rescale the different types of modalities, we adopted standardization. The standard scaler rescales all modal data to zero mean and unit variance.

C. Automatic Labeling and Segmentation

After preprocessing, we segmented the data into each single dance figure timestep. Figure 5 shows the segmentation method. As shown in the figure, we first calculate the number of samples in each dance figure from beat per minute(BPM) information and the number of figures in the performance. In the Ballroom Dance Dataset, the participants performed the choreography with the music of BPM 100 music. Moreover, every sequence contains 19 figures(comprising 13 figure types) in total, each of which has four beats. From this information and sampling rate, we calculate the number of samples in each dance figure as follows:

$$\begin{aligned} samples = (60[sec.] \times 120[Hz]) \\ \div (100[BPM] \div 4[beats\ per\ a\ figure]) = 288 \end{aligned} \tag{1}$$

where 60[sec.] × 120[Hz] calculates the number of samples in one minute, 100[BPM] ÷ 4[beats] calculates the number of

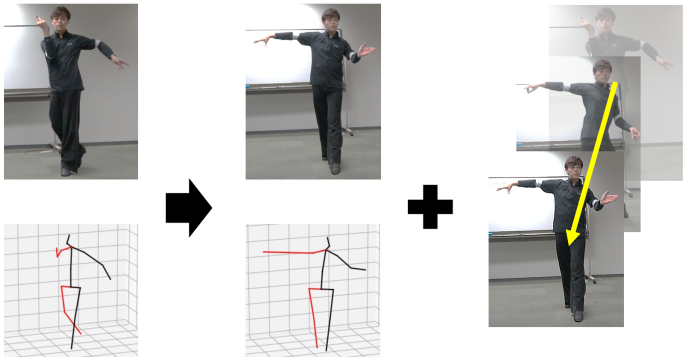


Fig. 4. Body coordinate processing and adding trajectory information.

sets of four beats (i.e., beats in a figure) in one minute. Thus, the formula provides the number of samples in each dance figure.

D. Temporal Masking Module

Given the segmented sequence of a dance figure, we apply a temporal masking mechanism. The objective of this mechanism is to increase the importance of actions in the middle of the time sequence. Generally, dance has a choreography of continuous dance figures, including ballroom dance. The continuous dance figures have the transition part of the two figures, changing the movement depending on the previous/following dance figure. Although these transition parts may contain information about each dance figure, the center part of the figure includes characteristics that are purer. Therefore, we must weigh the pure movement features. We achieved this objective by applying the one-dimensional Gaussian importance mask. The Gaussian mask works to suppress the importance of features when apart from the middle of the movement sequence (pure features) and close to the edges (transition points). The Gaussian mask was formulated as follows:

$$\tilde{X}_{t,c} = X_{t,c} \odot M^t \tag{2}$$

$$M_x^t = e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

where c is the channel (acceleration, angular velocity, 3D joints), t is the time, and therefore, $X_{t,c}$ in Equation 2 represents for an input tensor of a dance figure. M_x^t in Equation 3 formulates the Gaussian mask for one-dimensional data.

E. Network Structure and Training

The generated data are fed to the DNN architectures. As every ballroom dance figure has its time-sequential features of footstep combination and body movement, we adopted an LSTM-based structure. We inserted a dropout layer and performed batch normalization to prevent overfitting. We adopted ReLU as the activation function. The order of the dropout layers is provided the best performance after the preliminary experiment. In the experiment, we tried placing it before the

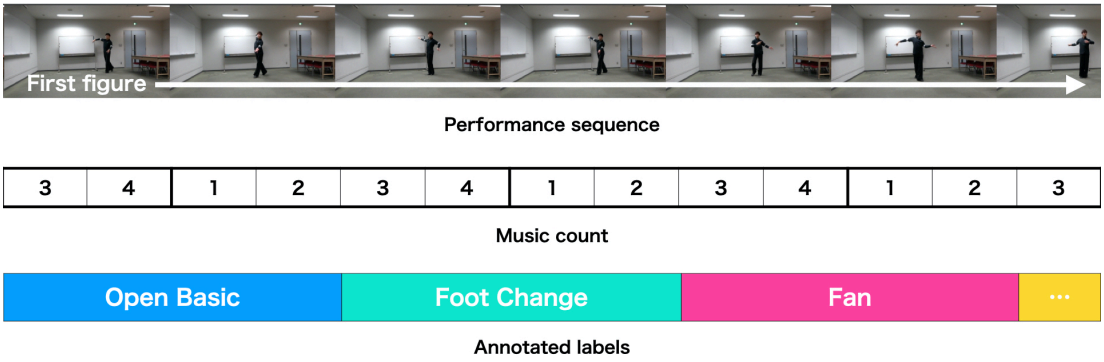


Fig. 5. Automatic dance figure labeling method.

LSTM layer, before the fully connected layer, and both. We found that placing it before the LSTM layer performs approximately 3% better than before the fully connected layer. In addition, placing both positions did not improve performance. Thus, we decided to place the dropout layer only before the LSTM layer. The batch normalization layer is fixed in its place after the LSTM layer and before the activation function to prevent the internal covariate shift of the LSTM outputs. Figure 2 shows an overview of the work process, including the DNN layer. Figure 3 shows our proposed LSTM-based DNN structure. As shown in Figure 3, the input tensor of every figure is fed to the model.

For wearable sensors, we provide a comparative verification of the positions to which a wearable sensor is attached. We prepared six parts: the right and left ankles, waists, and arms. We selectively added one of these modalities to the DNN classifier and compared the prediction accuracies. Finally, we chose one of the six body positions. Figure 6 shows the positions of the wearable sensors attached to the participants. In Latin American dance, each of the arms, hips, and legs moves characteristically. Therefore, we attached the sensors to all of them and compared the results. After the wearable sensors collected acceleration and angular velocity data, we preprocessed the data using standard scaling. Next, we prepared an LSTM layer independent of the LSTM of the 3D skeletal position and input sensor data (6 dim). Finally, the hidden status of the LSTM with 3D joints and wearable sensor data are concatenated to provide a final prediction.

The hyperparameters in the structure were selected using the grid-search algorithm. The learning rate was chosen from the 20 equally distributed points between 1e-3 and 1e-1. The hidden size was chosen from the eight equally distributed points between 32 and 256. The dropout rate is chosen from the nine equally distributed points between 0.1 and 0.9. The selected parameters were as follows: learning rate = 0.001, dropout rate = 0.2 (acceleration, angular velocity, and 3D joints), hidden size = 64 (acceleration, angular velocity), and 192 (3D joints).

V. Result

This section describes the classification results and discusses them in detail. First, we show the results of classification using 3D pose data. As our previous work used 2D pose data, we define the result with a 2D pose as our baseline. We compared and evaluated the results. Subsequently, we show the results supported by a wearable sensor, evaluating the best position on which the wearable sensor is attached. Next, we compared the F1-score of classification using 3D joints with and without trajectory information of mid-hip coordinates to show how our mid-hip trajectory information worked. (See Table IV for summary) For all results, we took the average of the trial-based five-fold cross-validation results. In the trial-based cross-validation, we controlled the training so that the segmented figures from the same performance recording would not appear in both training and validation. This is because we cannot use the figures from the same recording for training and testing in a real-world setting.

A. Human Testing

To compare our results with human accuracy, we recruited seven experienced dancers (not all of them were the same as the participants in the data collection phase). All of them had at least two years of experience in Latin American dance and knew all of the dance figure names in our dataset. First, each participant was asked to watch a video in which the segmented nine dance figure performances were randomly ordered and played. Duplication of figures is allowed in each video amid a participant inference using the elimination method. The number of dance figures that a participant watches is small compared to the training dataset for the model to reduce the burden of watching and annotating. Finally, we calculated the mean accuracy score for human annotations. The bottom row in Table IV shows the results.

B. Training and Validation

After a grid search, we set the batch size to 64, the optimizer as Adam, and the number of epochs to 300. We adopted trial-based metrics as the cross-validation methods. In the trial-based cross-validation, we split the dataset into training and validation so that the data from the same record did not appear in both training and validation. The trial-based metric assumes 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).

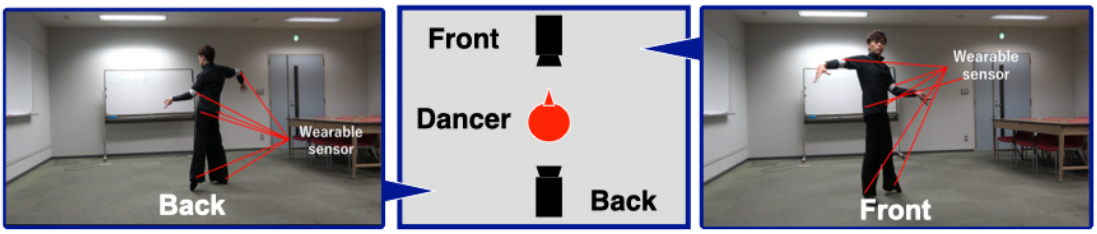


Fig. 6. Positions of worn sensors and shooting directions. There are two shooting directions: front and back. For each record, the video is shot from only one of them.

TABLE IV
SUMMARY OF THE CLASSIFICATION RESULTS.

Method	Accuracy
2D joints (baseline)	0.847
3D joints	0.876
3D joints + trajectory	0.896
3D joints + trajectory + wearable sensor	0.920
3D joints + trajectory + wearable sensor + temporal masking	**0.930**
Human (experienced dancers)	0.936

C. 3D Joints Result

Table V shows the classification results using 3D joints, and the baseline result with 2D joints in our latest work [11]. In our previous work [11], we utilized OpenPose [27]. However, in this study, we use VideoPose3D provided by FaceBook Research. Thus, we reproduce the 2D joint result using the algorithm of FaceBook Research as well. The figure column in Table V stands for each ballroom dance figure name. Precision, recall, and F1-score for each dance figure were provided. The last two rows at the bottom of the table show the overall accuracy and F1-score, respectively. Figure 7 shows a bar chart of F1-scores with 2D and 3D joints. The gray bars and orange bars stand for the F1-scores using 2D joints and 3D joints, respectively. The x-axis represents the dance-figure names, which are the labels to be predicted by our model. The y-axis, which stands for the F1-score, starts from 0.25 to 1.0 for easier understanding.

Table VI shows the comparative results of classification using 3D-joints with and without trajectory information. The figure column in Table VI represents each ballroom dance figure name. Precision, recall, and F1-score for each dance figure were given. The last two rows at the bottom of the table show the overall accuracy and F1-score, respectively. Figure 8 is a bar chart comparing the F1-score of dance figure classification with and without trajectory information of the mid-hip. The x-axis represents the dance-figure names, which are the labels to be predicted by our model. The y-axis, which stands for the F1-score, starts from 0.25 to 1.0 for easier understanding.

D. Results with Wearable Sensors

Table VII shows the results when wearable sensor data were added to the 3D joints. The column ”None” represents the results with 3D joints. “Left ankle,” “Left arm,” ..., “Right hip” stand for each position to which a wearable sensor is attached. The table shows the accuracy and F1-score as measurement tools of the result. Table VIII compares the number of sensors (one to six) attached to the body. The result for each number of sensors is the best score among all possible combinations of sensor positions.

E. Evaluation

We evaluated each experimental result of dance figure classification. We begin by evaluating the overall training and validation. Next, we compared 2D and 3D joint predictions, and then evaluated the effect of mid-hip trajectory information. Next, we verified the body positions to which a wearable sensor was attached, followed by the temporal masking module.

1) Overall Training and Validation: Figure 9 shows the training/validation loss and the validation F1-score while running train epochs, with 20% holding out. The epoch was set to 300. The training loss (blue) is smoothly decreasing, especially after 100 epochs. Meanwhile, the validation loss (yellow) is trembled while decreasing. In particular, the model overfits to the training data after 220 epochs. The model can include additional tuning to prevent this problem.

2) 2D and 3D Joints: Table V and Figure 7 compare the prediction results of the 2D and 3D joints. Overall, the accuracy and F1-score with 3D joints outperformed 2D joints by approximately 5%. However, 3D joint results do not outperform 2D joints in all figure classes. In “OpenBasic,” “HockyStick,” “OpeningOut,” and “Alemana,” the F1-score with 2D joints are better than 3D joints. In particular, the model with 3D joints found it difficult to recognize “HockyStick” and “Alemana”. Figure 10 compares the movements of HockyStick and Alemana. As a couple of people perform ballroom dance figures, some figures have almost the same movements for men at different figures. HockyStick and Alemana are good examples of this. However, there are some small differences in the body rotation and movement direction. The classification

TABLE V
COMPARISON OF CLASSIFICATION RESULTS USING 2D AND 3D JOINTS.

2D-joints				3D-joints			
Figure	Precision	Recall	F1-score	Figure	Precision	Recall	F1-score
OpenBasic	0.946	0.946	0.946	OpenBasic	0.961	0.875	0.916
FootChange	0.833	0.804	0.818	FootChange	0.806	0.964	0.878
Fan	0.810	0.839	0.825	Fan	0.855	0.839	0.847
HockyStick	0.895	0.607	0.723	HockyStick	0.550	0.786	0.647
NewYorkR	0.830	0.786	0.807	NewYorkR	0.839	0.929	0.881
NewYorkL	0.900	0.643	0.750	NewYorkL	1.000	0.929	0.963
SpotTurn	0.790	0.875	0.831	SpotTurn	0.915	0.964	0.939
NaturalTop	0.851	0.821	0.836	NaturalTop	0.920	0.821	0.868
OpeningOut	0.957	0.786	0.863	OpeningOut	0.885	0.821	0.852
Alemana	0.531	0.607	0.567	Alemana	0.571	0.286	0.381
HandToHandR	0.623	0.768	0.688	HandToHandR	0.764	0.750	0.757
HandToHandL	0.909	0.893	0.901	HandToHandL	0.945	0.929	0.937
Aida	0.813	0.929	0.867	Aida	1.000	0.893	0.943
Overall accuracy	0.847	Std	0.019	Overall accuracy	**0.896**	Std	0.016
Overall F1 score	0.839	Std	0.017	Overall F1 score	**0.893**	Std	0.017

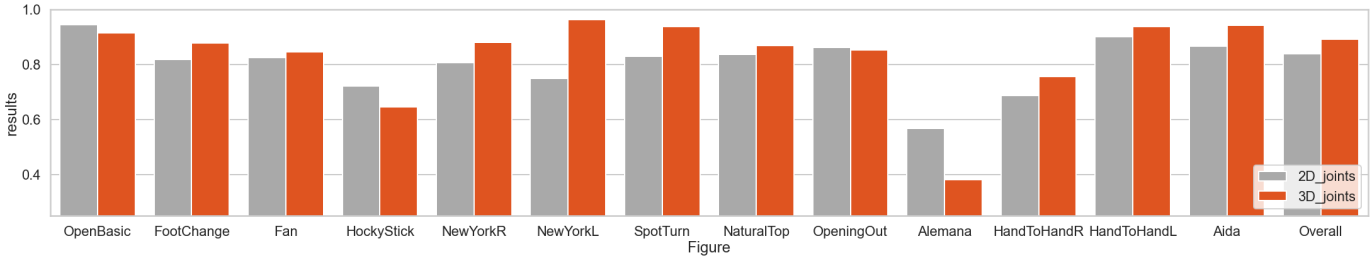


Fig. 7. Bar chart of Classification Results Using 2D and 3D joints.

TABLE VI
COMPARISON OF CLASSIFICATION RESULTS WITH AND WITHOUT TRAJECTORY INFORMATION.

3D-joints without trajectory				3D-joints with trajectory			
Figure	Precision	Recall	F1-score	Figure	Precision	Recall	F1-score
OpenBasic	0.946	0.929	0.937	OpenBasic	0.961	0.875	0.916
FootChange	0.852	0.929	0.889	FootChange	0.806	0.964	0.878
Fan	0.913	0.750	0.824	Fan	0.855	0.839	0.847
HockyStick	0.733	0.786	0.759	HockyStick	0.550	0.786	0.647
NewYorkR	0.831	0.875	0.852	NewYorkR	0.839	0.929	0.881
NewYorkL	0.733	0.786	0.759	NewYorkL	1.000	0.929	0.963
SpotTurn	0.911	0.732	0.812	SpotTurn	0.915	0.964	0.939
NaturalTop	0.875	0.750	0.808	NaturalTop	0.920	0.821	0.868
OpeningOut	0.833	0.893	0.862	OpeningOut	0.885	0.821	0.852
Alemana	0.593	0.571	0.582	Alemana	0.571	0.286	0.381
HandToHandR	0.662	0.804	0.726	HandToHandR	0.764	0.750	0.757
HandToHandL	0.912	0.929	0.920	HandToHandL	0.945	0.929	0.937
Aida	0.964	0.964	0.964	Aida	1.000	0.893	0.943
Overall accuracy	0.876	Std	0.021	Overall accuracy	**0.896**	Std	0.016
Overall F1 score	0.864	Std	0.018	Overall F1 score	**0.893**	Std	0.017

TABLE VII
RESULTS FOR EACH SENSOR POSITION.

position	None (3D)	Left ankle	Left arm	Left hip	Right ankle	Right arm	Right hip
Accuracy	0.896	0.919	**0.920**	0.916	0.915	0.913	0.918
F1-score	0.893	0.906	**0.910**	0.907	0.902	0.901	0.908

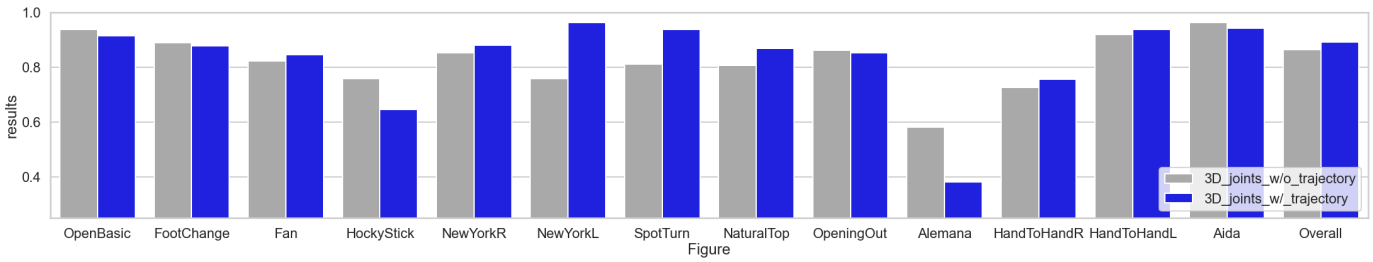


Fig. 8. Bar chart of Classification Results 3D Joints with and without Trajectory.

TABLE VIII
RESULTS WITH THE NUMBER OF ATTACHED SENSORS (1 - 6).

Number of wearable sensors (Best score among all possible combinations)	1	2	3	4	5	6
Accuracy	**0.920**	**0.920**	0.909	0.910	0.908	0.911
F1-score	**0.910**	0.908	0.900	0.904	0.898	0.901

TABLE IX
CLASSIFICATION RESULT WITH/WITHOUT TEMPORAL MASKING MODULE USING ACC, GYR, AND 3D JOINTS WITH TRAJECTORY.

Without temporal masking				With temporal masking			
Figure	precision	recall	F1-score	Figure	precision	recall	F1-score
OpenBasic	0.927	0.910	0.919	OpenBasic	0.964	0.946	0.955
FootChange	0.910	0.911	0.910	FootChange	0.944	0.910	0.927
Fan	0.943	0.893	0.917	Fan	0.941	0.857	0.897
HockyStick	0.758	0.785	0.771	HockyStick	0.774	0.857	0.813
NewYorkR	0.931	0.964	0.947	NewYorkR	0.965	0.982	0.973
NewYorkL	0.843	0.964	0.900	NewYorkL	1.000	0.964	0.982
SpotTurn	0.910	0.910	0.910	SpotTurn	0.890	0.875	0.882
NaturalTop	0.925	0.892	0.909	NaturalTop	0.840	0.750	0.792
OpeningOut	0.961	0.892	0.925	OpeningOut	0.889	0.857	0.873
Alemana	0.633	0.678	0.655	Alemana	0.731	0.679	0.704
HandToHandR	0.925	0.892	0.909	HandToHandR	0.911	0.911	0.911
HandToHandL	0.928	0.928	0.928	HandToHandL	0.915	0.964	0.939
Aida	0.892	0.892	0.892	Aida	0.900	0.964	0.931
Overall accuracy	0.920	Std	0.018	Overall accuracy	**0.930**	Std	0.015
Overall F1 score	0.911	Std	0.019	Overall F1 score	**0.922**	Std	0.017

results of 2D and 3D joints indicate that 2D joint information has some advantages in capturing these differences. Moreover, there is a suspicion that 2D joint information allows lets the classification model to learn the scene in which a dancer performs the two dance figures. The 2D joints provide information that depends on the video direction, whereas the 3D joints provide adequate information independent of the video direction.

3) Mid-hip Trajectory Information in a Timestep: Table VI and Figure 8 compare the prediction results using 3D joints with and without trajectory information. Overall, the accuracy and F1-score with trajectory information outperformed the result without trajectory information by approximately 3%. In particular, "NewYorkL," "SpotTurn," and "NaturalTop" achieved benefits from the trajectory information. From Figure 1, we can see that these three dance figures include circular movements in their footstep tracks. The mid-hip trajectory information helped the classification model understand the difference between figures that include body rotation but no circular footstep tracks and the figures that include body rotation and circular footstep tracks.

Meanwhile, the results with trajectories did not perform well in several dance figures. These figures have characteristics that their movements include body rotation of less than 90 $^{\circ}$, as shown in Figure 1. The results indicate that the mid-hip trajectory information is helpful for recognizing dance figures with significant body rotations and circular footstep tracks, while it confuses the model to handle the figures with a smaller amount of rotation.

4) Wearable Sensor Position and Number of Sensors: Table VII compares the prediction results obtained using 3D joints and functional wearable sensors. In terms of both in accuracy and F1-score, we obtained better results by adding wearable sensor data by approximately 3% on the average. Although the results among the six wearable sensor positions are close, the prediction using the wearable sensor of the left arm

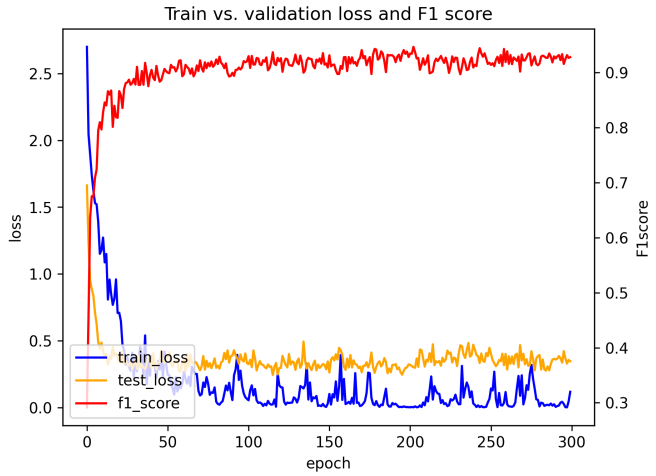


Fig. 9. Loss curve of training/validation and F1 score.

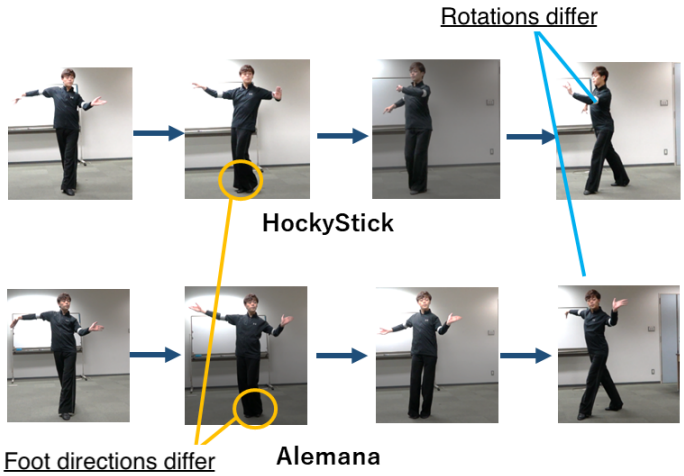


Fig. 10. Movements of HockyStick and Alemana.

outperformed the other parts. The left-arm information may be important to the classification model because the left arm's movement was different for each dance figure, guiding the partner. The left-arm action was also evident when the participants performed solo. Table VIII shows that there is almost no difference in the results when two or more sensors are added. Thus, we used only one wearable sensor to perform a dance-figure prediction.

5) Temporal Masking: Table IX compares the prediction results with and without applying the temporal masking module with the hybrid input of acceleration, angular velocity, and 3D joints with trajectory. With the temporal masking module, the overall accuracy and F1-score are improved by 1%, which is very close to the accuracy of experienced dancers. The temporal masking module applies a Gaussian mask to every dance-figure sequence. Figure 5 shows how the dance figures are connected to each other. Every two connected figures have their transition part, where the movement of the dance figures changes so that the figures are smoothly connected. The Gaussian mask gives low importance to those transition parts

TABLE X
ABLATION STUDY ON THE METHODS TO AVOID OVERFITTING.

Method	Accuracy
Dropout	0.926
Batch norm	0.891
Dropout + batch norm	**0.930**

and, contrastingly, provides high significance to the middle part of each dance figure, by which the pure characteristics of the figures are extracted.

F. Ablation Study

Our methodology includes dropout and batch normalization to avoid overfitting. However, it is reported that those two methods often work worse when used together [28]. Therefore, we explore how the results change when each or both of them are implemented. Table X presents the results. The results indicate that, in our study, using both of the methods provided the best performance. However, the use of dropout layers alone provided a close accuracy too.

VI. Future Work

Although our method showed a promising results in our ballroom dance dataset and human testing, the scalability of the method to the other activities was not evaluated. However, it is challenging to obtain an appropriate dataset to evaluate our technique. Our approach is for the multimodal dance activity data of acceleration, angular velocity, and video, where the camera setting and choreography are also carefully controlled. Therefore, we need to search for a suitable dataset for scalability evaluation.

Misrecognized figures can be analyzed in more interpretable ways. In the field of computer vision, the Grad-cam [29] is widely used to explore the image classification model. The Recurrent neural network has also been studied for interpretability [30]. Exploring the dance figure classification for explainability may reveal why our method found it difficult to classify some particular dance figures. Therefore, studying the dance figure classification model's explainability is one of the most exciting works following this paper.

VII. Conclusion

This study investigated a classification method for ballroom dance figures using 3D joints and a wearable sensor. We first evaluated the results using 3D joints by comparing them with 2D joints. Then, we investigated how the trajectory information of the middle hip coordinate improved the prediction. Subsequently, we compared the positions to which a wearable sensor was attached. In addition, we introduced a temporal masking module to extract the pure movements of the dance figures. As a result, we achieved 93% accuracy with our proposed method, which is highly overwhelming the baseline result (84.7%) and very close to the accuracy of the experienced dancers (93.6%). Our proposed method with the

trajectory and temporal masking module showed almost as high accuracy as that of experienced dancers'.

Although 3D joint data yielded a better result than 2D joints overall, there was a suspicion that 2D joint information allows the classification model to learn the scene in which a dancer performs the two dance figures. This information depends on the video direction and environment, so we need to investigate the problem by collecting the dance performance video shot from more directions. In addition, the mid-hip trajectory was not always useful, especially for the dance figures with smaller or no amount of body rotation. It is possible to consider a structure that uses trajectory in a better way to maintain accuracy when handling those dance figures with less rotation. The classification results can be analyzed using an interpretable machine learning method.

VIII. Acknowledgment

The ballroom dance performances were provided by 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. In *ACHI 2012 : The Fifth International Conference on Advances in Computer-Human Interactions*, 2012.

[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. In *The Virtual Reality Society of Japan*, 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] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[11] Hitoshi Matsuyama, Kei Hiroi, Katsuhiko Kaji, Takuro Yonezawa, and Nobuo Kawaguchi. *A Basic Study on Ballroom Dance Figure Classification with LSTM Using Multi-modal Sensor*, pages 209–226. Springer Singapore, Singapore, 2021.

[12] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] K. Wang, J. He, and L. Zhang. Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors. *IEEE Sensors Journal*, 19(17):7598–7604, 2019.

[14] Q. Teng, K. Wang, L. Zhang, and J. He. The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition. *IEEE Sensors Journal*, 20(13):7265–7274, 2020.

[15] S. M. Mathews, C. Kambhamettu, and K. E. Barner. Centralized class specific dictionary learning for wearable sensors based physical activity recognition. In *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, 2017.

[16] Sherin M. Mathews, Chandra Kambhamettu, and Kenneth E. Barner. Maximum correntropy based dictionary learning framework for physical activity recognition using wearable sensors. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Fatih Porikli, Sandra Skaff, Alireza Entezari, Jianyuan Min, Daisuke Iwai, Amela Sadagic, Carlos Scheidegger, and Tobias Isenberg, editors, *Advances in Visual Computing*, pages 123–132, Cham, 2016. Springer International Publishing.

[17] Sherin M. Mathews. Dictionary and deep learning algorithms with applications to remote health monitoring systems. 2017.

[18] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *CoRR*, abs/1706.08033, 2017.

[19] Arshad Jamal, Vinay P. Namboodiri, Dipti Deodhare, and K. Venkatesh. Deep domain adaptation in action space. In *BMVC*, 2018.

[20] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib. Human action recognition using transfer learning with deep representations. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 463–469, 2017.

[21] Paradiso Joseph, Hu Eric, and Hsiao Kai yuh. The CyberShoe: A Wireless Multisensor Interface for a Dancers Feet. pages 57–60, 03 1999.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Xiang Li, Shuo Chen, Xiaolin Hu, and Jian Yang. Understanding the disharmony between dropout and batch normalization by variance shift, 2018.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[30] A. Karpathy, J. Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *ArXiv*, abs/1506.02078, 2015.