Journal of Multimedia Information System VOL. 11, NO. 2, June 2024 (pp. 149-156): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2024.11.2.149

Multi-Modal Beat Alignment Transformers for Dance Quality Assessment Framework

Taewan Kim[1*]

Abstract

In recent years, the dance entertainment industry has been growing as consumers have a desire to learn and improve their dancing skills. To fulfill this need, they need to be evaluated and get feedback to improve their dance skills, but the evaluation is very dependent on professional dancers. With the advent of deep learning techniques that can understand and learn the structure of 3D skeletons, graph convolutional networks and transformers have shown performance improvements in 3D human action understanding. In this paper, we propose Dance Quality Assessment (DanceQA) Framework to evaluate dance performance and predicts its dance quality numerically. For problem definition, we collect and capture 3D skeletal data by 3D pose estimator and label their dance quality. By analyzing the dataset, we propose dance quality measures, kinematic information entropy and multi-modal beat similarity, which consider traditional criteria for dance techniques. Based on results of the dance quality measures, kinematic entropy embedding matrix and multi-modal beat alignment transformers are designed to learns salient joints and frames in 3D dance sequence. Thus, we design the overall network architecture, DanceQA transformers, which consider spatial and temporal characteristics of 3D dance sequence from multiple input features and demonstrate that the proposed transformers outperform other Graph Convolutional Network (GCN)s and transformers on the DanceQA dataset. In numerous experiments, the CQTs outperforms previous methods, graph convolutional networks and multimodal transformers, at least by up to 0.146 in correlation coefficient.

Key Words: Dance Quality Assessment, Multi-Modal Learning, Kinematic Entropy, Transformers.

I. INTRODUCTION

Dance Quality Assessment (DanceQA) evaluates techniques of performers dancing to music and provides their dance performance quality numerically. Along with people's interest in dancing, the size of the industry market related to dancing has been growing over recent years [1], and a large number of dance videos have been uploaded on social media. In addition, Breaking, a style of dance that originated in the United States in the 1970s, has been chosen to feature on the Paris 2024 Olympic sports program as a new sport [2]. In response to this general interest, the criteria for dance performance evaluation have evolved specifically. Researchers have studied guides for dance performance evaluation by judges considering physical ability and rhythmic accuracy of the performers [3-5]. However, the hurdle of specialized knowledge makes non-experts inaccessible to DanceQA, and this leads to the need for DanceQA algorithm which can learn the knowledge and evaluate automatically.

The automation of DanceQA saves evaluators' efforts by providing score predictions of dance performance quality. When evaluating dance performance, the human evaluators must watch every single video and determine performance rating while closely looking into dance motions. Without the help of automated DanceQA, the dance performance evaluation requires a significant amount of time for the evaluators to watch a lot of dance videos which are few minutes long. The automated DanceQA provides dance performance scores and helps to select candidates which the evaluators should watch for sophisticated evaluation. This secondary role not only saves human resources, but also allows for more efficient assessments by focusing the expert's efforts on critical parts.

The automated DanceQA is able to fulfill a key role as a main evaluator in situations where an expert is not available. Since existing dance performance evaluation methods [3-4] are very dependent on the dance experts, they are rarely used in everyday situations due to their low accessibility. It is difficult for non-experts to utilize DanceQA, which requires a lot of time and effort to learn related expertise for accurate evaluation. The automated DanceQA method can replace this process by training correlation between dance performance and its quality score with high

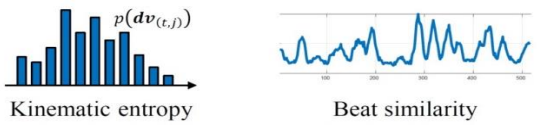
Manuscript received May 20, 2024; Revised June 05, 2024; Accepted June 19, 2024. (ID No. JMIS-24M-05-019)
Corresponding Author (*): Taewan Kim, +82-940-4751, kimtwan21@dongduk.ac.kr
[1]Division of Future Convergence (Data Science Major), Dongduk Women's University, Seoul, Korea, kimtwan21@dongduk.ac.kr

availability.

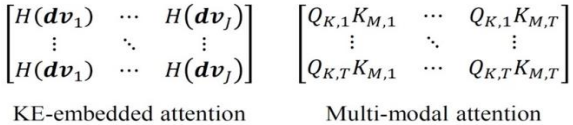
The automated DanceQA provides analysis of dance performance at joint or frame level to give feedback to a performer. In dance performance evaluation, the human evaluators usually give comments on the performance to the performer. Since the human evaluators are not always available, previous works [6-8] have analyzed dance performance only the kinematic data of dance motion like joint positions or joint angles. This information is only shown as a graph and is not visualized for human perspective, making it difficult to understand intuitively. Deep learning techniques which try to capture neural activation [9] or an attention weight [10] have been developed to analyze prediction results visually. DanceQA networks can visualize the activation of spatial joints or temporal frames which contribute to the dance quality score by utilizing these techniques.

In this paper, we propose Dance Quality Assessment Framework to handle issues and include datasets, dance quality measures, and regression networks for DanceQA as shown in Fig. 1. 3D motion data are collected in two ways, 3D pose estimation and motion capture, to build DanceQA dataset for subjective test which provides a dance quality score by ranking performers from relative comparison results. For dance quality labeling, Performance Competence Evaluation Measure (PCEM) [11] guides are adopted, and the subjective test involves comparing a pair of dance performance and choosing the relatively better one following the guides. To capture important elements of dance motion,

(a) Dance and music data.



(b) Dance quality measure.



(c) DanceQA transformers.

Fig. 1. Main components for dance quality assessment framework. The proposed framework uses multi-modal data, (a) dance and music data, and (b) dance quality measure is designed to examine the statistics of each joint and the multimodal beat alignment. Based on the results of these measures, the core modules in (c) DanceQA transformers are proposed to predict dance quality scores accurately.

Kinematic Information Measure and Kinematic-Music Beat Alignment are designed by examining kinematic statistics and multimodal similarity. To target the dance quality score ranked by the subjective test, the regression networks sufficiently long and dense 3D motion by considering diverse characteristics of 3D skeletal sequence and fusing them in transformers while referring results of the dance performance measures. Lastly, to measure kinematic and music beat alignment, multi-modal attention blocks are designed to train correlation between multi-modal inputs, dance motion and music.

At first, 3D motion data are inferred by 3D pose estimator from RGB videos, which are captured by the public and uploaded to the web. These videos have diversity in terms of dance quality, but the inferred 3D motion is not accurate due to limitations of the estimator. By leveraging kinematic characteristics of the 3D dance motion, dance proficiency is measured by the kinematic information entropy and the multi-modal similarity based on previous quality assessment methods [12]. A new representation of 3D motion is proposed to link spatially or temporally adjacent joints according to natural connection of skeletal structure. Existing spatial or temporal difference vectors [10] are replaced by the proposed dependency inputs, which show the higher correlation between subjective scores and score predictions in experimental results. To regress the dance quality scores from 3D motion data, we employ transformers [13] as a baseline, which trains spatial or temporal dependency via self-attention matrices. After processing each input, two streams are fused in fusion transformers instead of ensemble learning which is used commonly in human action understanding. Multi-modal transformers learn correlation between the trained dance motion features and music features for beat alignment and the output features are regressed by the subjective quality scores.

In summary, this research presents a novel approach to dance performance evaluation by integrating kinematic information entropy and multi-modal beat similarity. In section II, we propose the new dance quality transformers model, and we will show the performance through many experiments in Section III. The originality lies in leveraging deep learning to automate and enhance the objectivity of dance quality assessment. The main contributions of the proposed DanceQA framework can be included as follows:

- Dance performance measures are proposed to leverage kinematic characteristics and beat alignment for examining important factors in DanceQA.
- A new representation of 3D skeletal motion is proposed to link spatially or temporally adjacent joints for natural connection of skeletal joints.
- Intra-motion transformers are designed by embedding the kinematic entropy to capture the dance quality in

spatial dimension and fuse the features with different characteristics. The dance motion features, and music feature are trained together for multi-modal learning by the proposed inter-motion and multi-modal transformers.

II. DANCE QUALITY TRANSFORMERS

Overall network architecture is shown in Fig. 2. The dataset including 3D dance motion data and its label for DanceQA is stored in the database of Fig. 2 and 3D skeletal data are represented by the features with different characteristics and used as inputs to these networks. Intra-motion Transformers learn spatial dependency of each feature within the spatial tokens based on entropy-embedded attention, and these features are fused by Fusion Spatial Transformers to consider diverse characteristics of features. Temporal dependency is trained with following Inter-motion Transformers, and music features are trained together in multi-modal Transformers to align kinematic and musical beats. The details of the proposed networks for 3D DanceQA will be described in following subsections.

2.1. Understanding of Choreography

The Dance Quality Assessment predicts dance quality score from dance performance in terms of 3D human motion. The 3D human motion data consists of 3D skeletal sequence that articulates human motion with major joint positions. Let p be joint positions of the skeletal sequence and p is defined as follows:

$$p = \{p_{(t,j)} \in \mathbb{R}^3 | t = 1, \dots, T, j = 1, \dots, J\}, \quad (1)$$

where T and J is the number of frames and joints of the 3D skeleton sequence. The dance quality score y is labeled and coupled with its corresponding 3D skeleton sequence p. The DanceQA dataset can be defined as follows:

$$D = \{(p^{(1)}, y^{(1)}), (p^{(2)}, y^{(2)}), \dots, (p^{(N)}, y^{(N)})\}, \quad (2)$$

where N is the number of data pairs in the dataset.

2.2. Feature Representation

Previous works usually used these joint positions or their spatial and temporal difference vectors, which represent bone and velocity vectors, to learn semantics of 3D skeleton sequence. In DanceQA, the body shape and movement in dance performance can be articulated by these two features, which are very useful to predict its dance quality score. These features are also used in the proposed network to consider diverse characteristics of dance motion. However, these features could not represent real bone and velocity vectors because their start and end positions of the vectors disappeared as shown in Fig. 3(a). They are just vectors starting from the origin, and learning algorithms cannot understand the bone and velocity information the way human perceives.

To help networks remember the information, as shown in Fig. 3(b), we propose joint-dependent features which con-

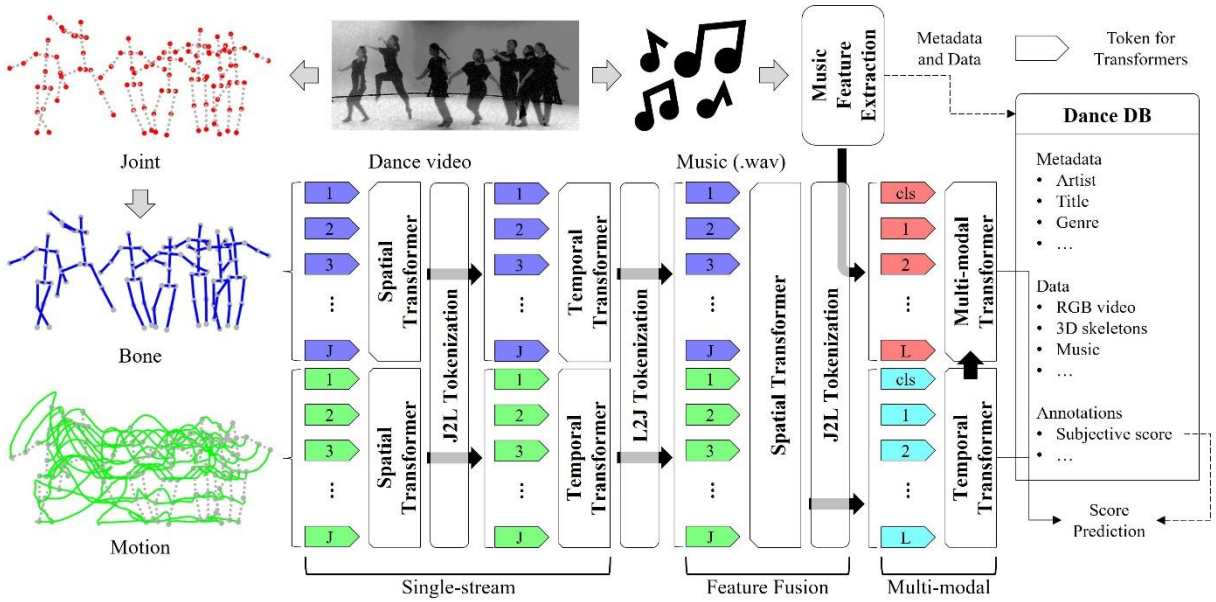
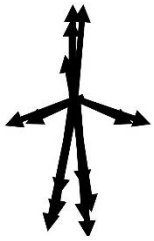
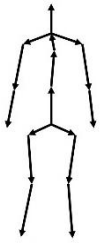


Fig. 2. Data preprocessing and DanceQA network architecture. 3D skeletons and music features are extracted from the dance video. Dance videos, 3D motion data, and their annotations are stored in the database with their metadata which describes dance performance. The skeletal data is represented as joint, bone, and motion features, which are used as inputs for DanceQA networks. These networks consist of Spatial, Temporal, and Multi-modal Transformers and predict a dance quality score of the input 3D sequence.



(a) Difference vector representation



(b) Joint dependency representation

Fig. 3. Bone features represented in the form of (a) difference vector representation and (b) joint dependency representation respectively.

tain the start and end positions of the bone and velocity information instead of spatial and temporal difference. This is more similar to human perception than the difference vector representation when observing the skeleton sequence. The bone dependent (spatially joint-dependent) features b are represented as follows:

$$b = \{b_{(t,j)} \in \mathbb{R}^6 | p_{(t,i_1)} \oplus p_{(t,i_2)}, i = 1, \dots, J-1\}, \quad (3)$$

where $\oplus$, i_1, i_2 is the operation to concatenate vectors, the start and end joints of i^{th} bone, respectively. The motion dependent (temporally joint-dependent) features m are represented as follows:

$$m = \{m_{(t,j)} \in \mathbb{R}^6 | p_{(t-1,j)} \oplus p_{(t,j)}, t = 2, \dots, T\}. \quad (4)$$

The spatial and temporal dependency of skeletons is embedded in these features respectively while preserving joint positions. In transformers, it is difficult to reflect the joint relationship without any constraint on spatial or temporal self-attention matrices. Nevertheless, the dependent features make it possible to train the bone and motion vectors simply by connecting dependent joints in preprocessing step without information loss and improve dance quality prediction performance in DanceQA.

2.3. Input Encoding

Previous Skeleton has the characteristics of both natural language and images in spatial and temporal dimensions, respectively. The joints have specific semantics such as neck, hip, shoulder, elbow, wrist, knee, ankle and so on like words in natural language, but the frames are not semantically separated like pixels in images. Thus, each joint is used as a spatial token and multiple frames are used as a temporal token to explore spatial or temporal dependency in transformers.

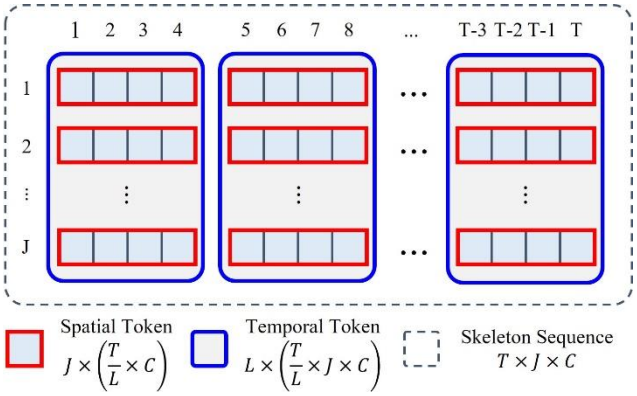


Fig. 4. Spatial and temporal tokens for the transformers. In the Spatial Transformers, the spatial dependency of joints within the basic motion is trained locally, and the temporal dependency of the temporal tokens is trained globally to predict the dance quality score.

The spatial and temporal tokens are described as the red box and the blue box, respectively, in Fig. 4. The input features are reshaped into the spatial token, which is the shortest unit to analyze long dance sequence. The i^{th} spatial token of the j^{th} joint $X_{S,(l,j)}$ is defined as follows:

$$X_{S,(l,j)} = \{X_{\left(\frac{T}{L}\times(l-1)+1,j\right)}, \dots, X_{\left(\frac{T}{L}\times l,j\right)}\} \in \mathbb{R}^{\frac{3T}{L}}, \quad (5)$$

where L is the number of tokens. The spatial tokens are concatenated as the temporal token including all the joints. The l^{th} temporal token is defined as follows:

$$X_{T,(l)} = \{X_{S,(l,1)}, \dots, X_{(S,(l,J))}\} \in \mathbb{R}^{\frac{3TJ}{L}}. \quad (6)$$

Each spatial token is embedded with the channel size CT/L for the spatial transformers, and the temporal tokens are used as inputs for the temporal transformers after J2L Tokenization where L is the number of the temporal tokens. The number of frames within the token T/L is very important factor for DanceQA. The temporal token with the small number of frames cannot represent sufficiently the unit motion while it is difficult to understand the dance sequence if the unit motion is too long. We find the optimal T/l empirically by testing multiple values in next Section.

2.4. Intra- and Inter-Motion Transformers

The Intra-motion transformers are designed to train not only spatial dependency of diverse skeletal features but also short-term characteristics of the unit motions. After the input embedding, the positional embedding is injected for the

transformers to make use of the joint positions. For the inter-motion transformers, outputs of the intra-motion transformers are concatenated into the temporal token. As shown in Fig. 4, all the spatial tokens within the unit motion are gathered together for the temporal token.

After the intra-motion and inter-motion transformers, the correlation between kinematic and musical beats are trained by the multi-modal transformers in Fig. 2. The skeletal features are trained by self-attention of the inter-motion transformers while the music feature is trained by cross-attention of the multi-modal transformers.

2.5. Overall Architecture Design

The proposed overall network architecture is proposed as shown in Fig. 2 consisting of the input features and transformers. Single-Stream Transformers (SST) learn spatial and temporal dependency of each feature. Feature Fusion Transformers (FFT) learn dependency between multiple features, which are trained sufficiently from each SST. Understanding each feature independently helps to fuse multiple features together. At last, for multi-modal learning between kinematic and music features, multi-modal transformers and inter-motion transformers are trained together. The inter-motion transformers learn temporal dependency of the outputs of the FFT only by their self-attention. The multi-modal transformers learn correlation between the musical features extracted from wave files and the kinematic features from the inter-motion transformers. The quality tokens denoted as Q of multi-modal beat alignment and inter-motion transformers in Fig. 2 are regressed by the linear layer to predict the dance quality score. Mean Squared Error (MSE) is employed as loss function to train the proposed transformers by decreasing the difference between dance quality prediction and its subjective score.

III. EXPERIMENTS

3.1. Dataset Protocol

To measure performance of the proposed dance quality measure and DanceQA networks, we suggest two protocols. Protocol I contains only 'Dynamite' of BTS. 23 videos are used for training and 10 videos are used for testing. Protocol II contains only 'Kill this love' of BLACKPINK. 21 videos are used for training and 10 videos are used for testing. Protocol I and II are tests to measure the quality of dance performance for the same choreography.

3.2. Performance Comparison

To show the performance of the proposed transformers, ST- Graph Convolutional Network (GCN) [14], AGCN [10] and FACT [15] are tested with various features. ST-GCN and AGCN are the most popular graph convolutional networks in action recognition. These networks build adjacency matrices according to the natural connection of skeleton and convolve neighbored joints. FACT uses full attention to find kinematic and musical relationship at frame level without considering spatial structure within a frame.

For performance comparison, we used two metrics: Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Correlation Coefficient (SRCC). PLCC measures the linear relationship between two continuous variables. It provides a value between −1 and 1, where 1 indicates a perfect positive linear correlation, −1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation. Moreover, SRCC measures the strength and direction of the monotonic relationship between two ranked variables. It is a non-parametric measure and provides a value between −1 and 1, similar to PLCC.

FACT and ST-GCN shows the lowest and the highest correlation, respectively, among the other methods. The tokens of FACT include both 3D skeletons and musical features, which blend and seem to disturb predicting their dance quality score, and this leads to the design of the multi-modal transformers that keep the kinematic feature from the musical feature. ST-GCN and AGCN have their adjacency matrices, which define connections among joints or bones. This definition may be very good constraint on action recognition but does not help to find salient joints in the DanceQA. There are salient joints, especially an ankle in this case, which have high correlation with the dance quality score. This distinct point leads to using transformer blocks to identify salient joints readily for dance quality prediction.

3.3. Component Analysis

For ablation study, three types of networks are listed in Table 1 with various input features. Inter-motion only considers the temporal relationship between the tokens. Intra-motion+Inter-motion handles both spatial and temporal attentions sequentially and show performance improvement compared to Inter-motion only. Intra-motion+Inter-motion +Multi-modal BA covers cross-modal correlation between kinematic and musical features beyond spatial and temporal modeling and show the significance of musical information in the DanceQA. Intra-motion+Inter-motion+Multi-modal BA+KE embedding adds kinematic entropy embedding to attention matrices in intra-motion transformers.

There are the performance improvements with additional transformers or musical features. Inter-motion only shows the lowest accuracy in the ablation study, and the addition of intra-motion transformers makes focused on the salient joints. As previously mentioned about FACT, it is very critical for the quality regressors to figure out spatial charac-

Table 1. Results of proposed method.

Networks	Feature	Dynamite		Kill this love	
		PLCC	SRCC	PLCC	SRCC
ST-GCN	Joint	0.6838	0.5857	0.7111	0.5982
	Bone	0.7009	0.5920	0.7351	0.6056
	Motion	0.6966	0.5808	0.7298	0.5988
AGCN	Joint	0.6789	0.5856	0.6927	0.6103
	Bone	0.6860	0.5785	0.7070	0.6061
	Motion	0.6858	0.6028	0.7080	0.6219
Multimodal transformer	Joint+music	0.6456	0.5393	0.6722	0.5587
	Bone+music	0.6326	0.5263	0.6477	0.5466
	Motion+music	0.6303	0.5165	0.6487	0.5315
Inter-motion only	Joint	0.7494	0.6850	0.7769	0.6994
	Bone	0.7396	0.6960	0.7696	0.7171
	Motion	0.7260	0.6772	0.7407	0.7030
Intra-motion+inter-motion	Joint	0.8042	0.7151	0.8410	0.7389
	Bone	0.8202	0.7464	0.8573	0.7660
	Motion	0.8025	0.7179	0.8249	0.7381
	Bone+motion	0.8231	0.7651	0.8313	0.7910
Intra-motion+inter-motion+multimodal BA	Joint+music	0.8166	0.7611	0.8497	0.7792
	Bone+music	0.8160	0.7642	0.8547	0.7891
	Motion+music	0.8028	0.7419	0.8226	0.7690
	Bone+motion+music	0.8221	0.7770	0.8609	0.8028
Intra-motion+inter-motion+multimodal BA+KE embedding	Bone+motion+music	0.8310	0.7852	0.8654	0.8123

teristics for dance quality prediction. Also, cross-attention of multi-modal transformers improves the prediction performance by learning cross-correlation between kinematic and musical features regardless of the feature type.

IV. CONCLUSION

In this paper, we present a DanceQA framework based on transformer architecture. The proposed dance quality measures, kinematic entropy, and multi-modal beat similarity, remarkably distinguish salient joints in the DanceQA. The proposed DanceQA transformers capture salient body parts and frames in their attention weights that contribute to the dance quality prediction while outperforming other GCNs and transformers. We present various experimental results and analyses which show how to capture the dance quality. This framework demonstrates the feasibility of the DanceQA in 3D skeleton domain for future works. We train only single choreography, but the DanceQA framework needs to handle more than two choreographies for evaluating unseen choreography.

ACKNOWLEDGMENT

This study was supported by the Dongduk Women's University grant.

REFERENCES

[1] A2Z Market Research, Online Dance Training Market Recovery and Impact Analysis Report Steezy Studio, DancePlug, Dancio.

[2] https://www.paris2024.org/en/sport/breaking/

[3] S. J. Chatfield and W. C. Byrnes, "Correlational analysis of aesthetic competency, skill acquisition and physiologic capabilities of modern dancers," in *5th Hong Kong International Dance Conference Papers.* Hong Kong: The Secretariat of the Hong Kong Academy for the Performing Arts, 1990, pp. 79-100.

[4] A. A. Parrott, "The effects of Pilates technique and aerobic conditioning on dancers' technique and aesthetic," *Kinesiol Med Dance*, vol. 15, no. 2, pp. 45-64, 1993.

[5] Y. Koutedakis, H. Hukam, G. Metsios, A. Nevill, G.

Giakas, and A. Jamurtas, et al., "The effects of three months of aerobic and strength training on selected performance-and fitness-related parameters in modern dance students," *The Journal of Strength & Conditioning Research*, vol. 21, no. 3, pp. 808-812, 2007.

[6] J. C. Chan, H. Leung, J. K. Tang, and T. Komura, "A virtual reality dance training system using motion capture technology," *IEEE Transactions on Learning Technologies*, vol. 4, no. 2, pp. 187-195, 2010.

[7] S. Laraba and J. Tilmanne, "Dance performance evaluation using hidden Markov models," *Computer Animation and Virtual Worlds,* vol. 27, no. 3-4, pp. 321-329, 2016.

[8] Y. Kim and D. Kim, "Interactive dance performance evaluation using timing and accuracy similarity," in *ACM SIGGRAPH 2018 Posters*, 2018, pp. 1-2.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921-2929.

[10] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition,* 2019, pp. 12018-12027.

[11] D. Krasnow and S. J. Chatfield, "Development of the performance competence evaluation measure: Assessing qualitative aspects of dance performance," *Journal of Dance Medicine & Science*, vol. 13, no. 4, pp. 101-107, 2009.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. N. Gomez, et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 32, no. 1, 2018.

[15] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13401-13412.

AUTHOR

Taewan Kim received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2008, 2010, and 2015, respectively. From 2015 to 2021, he was with the Vision AI Laboratory, SK Telecom, Seoul. In 2022, he joined as a faculty with the Division of Future Convergence (Data Science Major), Dongduk Women's University, Seoul, where he is currently an Assistant Professor. His research interests include computer vision and machine learning including continual and online learning.