International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Beyond the Screen With DanceSculpt: A 3D Dancer Reconstruction and Tracking System for Learning Dance

Sanghyub Lee, Woojin Kang, Jin-Hyuk Hong & Duk-Jo Kong

To cite this article: Sanghyub Lee, Woojin Kang, Jin-Hyuk Hong & Duk-Jo Kong (2025) Beyond the Screen With DanceSculpt: A 3D Dancer Reconstruction and Tracking System for Learning Dance, International Journal of Human–Computer Interaction, 41:9, 5406-5419, DOI: 10.1080/10447318.2024.2360773

To link to this article: https://doi.org/10.1080/10447318.2024.2360773

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
2025, VOL. 41, NO. 9, 5406–5419
https://doi.org/10.1080/10447318.2024.2360773

Beyond the Screen With DanceSculpt: A 3D Dancer Reconstruction and Tracking System for Learning Dance

Sanghyub Lee[a*], Woojin Kang[a*], Jin-Hyuk Hong[a,b], and Duk-Jo Kong[b,c]

[a]School of Integrated Technology, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea; [b]AI Graduate School, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea; [c]Center for Research Innovation, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea

ABSTRACT

Dance learning through online videos has gained popularity, but it presents challenges in providing comprehensive information and personalized feedback. This paper introduces DanceSculpt, a system that utilizes 3D human reconstruction and tracking technology to enhance the dance learning experience. DanceSculpt consists of a dancer viewer that reconstructs dancers in video into 3D avatars and a dance feedback tool that analyzes and compares the user's performance with that of the reference dancer. We conducted a comparative study to investigate the effectiveness of DanceSculpt against conventional video-based learning. Participants' dance performances were evaluated using a motion comparison algorithm that measured the temporal and spatial deviation between the users' and reference dancers' movements in terms of pose, trajectory, formation, and timing accuracy. Additionally, user experience was assessed through questionnaires and interviews, focusing on aspects such as effectiveness, usefulness, and satisfaction with the system. The results showed that participants using DanceSculpt achieved significant improvements in dance performance compared to those using conventional methods. Furthermore, the participants rated DanceSculpt highly in terms of effectiveness (avg. 4.27) and usefulness (avg. 4.17) for learning dance. The DanceSculpt system demonstrates the potential of leveraging 3D human reconstruction and tracking technology to provide a more informative and interactive dance learning experience. By offering detailed visual information, multiple viewpoints, and quantitative performance feedback, DanceSculpt addresses the limitations of traditional video-based learning and supports learners in effectively analyzing and improving their dance skills.

KEYWORDS

Human computer interaction; human-centered computing; dance support tool; dance learning; 3D human reconstruction

1. Introduction

Dance is a form of art and communication that is loved by millions of people around the world (Zhou et al., 2021), and its interest continues to grow, especially with online platforms facilitate the continued interest. People have easy access to online dance videos, and often use them for educational purposes (Hong et al., 2020). Online platforms such as YouTube offer tons of music videos and dance tutorials for free. This has created an environment where people can learn and practice dance by their own without being constrained by space and time, and new cultural and social connection has been built through the sharing of dance. The phenomenon of sharing dance videos in so-called "dance challenges" on TikTok and other platforms is a prime example of building these social connections (Ng et al., 2021). However, this video-based self-learning has some challenges that the videos do not comprehensively show all the information, and provide feedback tailored to a dance learner's specific needs (Hou, 2022). People have a need to get more useful dance information to learn to dance better.

For this purpose, we first explored more specific needs of dance learners who often watch online dance videos. A preliminary survey was carried out with 21 students who were active in dance clubs. They had an average of 2.4 years ($SD = 1.2$) of dance practice and experience. According to the survey, when practicing dance, they mainly search for and watch many dance videos using their laptops and smartphones. They heavily relied on dance videos to get most of dance information such as postures and movements of reference dancers, and then practiced in front of a mirror to check their dancing. They repeated this process until they've learned enough about the dance. They also filmed their dancing and compared it with the reference dancers' dancing in the videos for self-evaluation (Lee & Lee, 2023). What they were commonly interested in is dance videos that include original choreography to a music (e.g., hip-hop, K-pop) or choreography that has been recreated in different styles (e.g., street dance, ballet). Since many videos highlight

CONTACT Duk-Jo Kong dukjokong@gist.ac.kr AI Graduate School, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea; Center for Research Innovation, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea
*Authors contributed equally to the work

certain parts of a dancer (face, hands, etc.), as shown in Figure 1(a), this often limits the understanding of the entire posture. Obviously, our survey participants preferred videos where the camera view is fixed and shows the dancer's entire body, as shown in Figure 1(b).

When watching dance videos, people observe reference dancers and try to mimic their dance pose and movement. Not all the videos are filmed to teach the dance (only a few actually are), and they're often disturbed by something blocking the dancer in the camera's view. Especially in synchronized dance where there are many dancers, it is very difficult to keep track of each dancer's movements and positions, as they frequently block each other and move in complex trajectories and formations. By watching and comparing many different videos of the same dance, the learners try to understand dance moves and formations deeply. But in practice, it's not easy to collect the same dance footage from as many different camera angles and distances as they want. Comparing their dance with the reference dancer in the video is another challenging issue. Learners can't easily tell which elements of their dance (e.g., pose, trajectory, formation, and timing) need to be modified just by watching theirs and reference videos. How much harder would it be for a beginner? To address these challenges, in this paper, we propose a DanceSculpt (DS) system that utilizes 3D human reconstruction & tracking (3D HRT) technology.

Our DS system provides (1) a dancer viewer that utilizes artificial intelligence (AI) to reconstruct a dancer in a video into a 3D avatar; and (2) a dance feedback panel that shows dance performance in terms of pose, trajectory, formation, and timing by comparing the choreography of 3D avatars of the reference dancer and a user. We investigated how people learn to dance in comparison of our DS system and common dance videos. (We call this video-watching approach as Conventional Practice (CP).) The comparative evaluation has been done by examining participants' behavioral responses and surveys, and by analyzing their dance performance.

In summary, this work has two main contributions:

- C1. We identified how people learn new choreography from dance videos, and designed 3D dancer visualization and performance analysis to meet their needs.
- C2. We developed DS, a dance learning system that addresses the limitation of typical dance videos by exploiting 3D HRT.

2. Background

2.1. Dance learning aids

The educational aspects of dance present complex challenges. Beginners often used to train under the guidance of an expert to learn elements such as rhythm, posture, and expression. In typical dance training, dance professionals usually show sample dance videos and demonstrate moves by themselves. They also analyze learners' dance performance, and give some verbal feedback or directly correct their pose or movements. For professional dancers, there are some analytical systems, such as Laban Movement Analysis (Whittier, 2006), Benesh Movement Notation (Singh et al., 1983), and Eshkol-Wachmann Notation (Guest, 1990), to communicate technical dance knowledge such as movement sequences and trajectories. These analytical systems facilitate understanding and communicating complex dance techniques. However, the use of these specialized analytical systems is limited because they are tailored to specific genres of dance and require expert training. To complement these traditional learning methods, a variety of learning aids and educational technologies have emerged. This includes various type of instructional videos (Sukel et al., 2003) and specific system that utilize motion capture in the environment of virtual reality (VR) (Aristidou et al., 2015).

Recently, the advances in the techniques of computer vision have been developed to analyze and evaluate the poses of dancers in images and videos. In particular, pose information, usually expressed as a skeleton structure, was mainly selected and used for comparative evaluation (Tsuchida et al., 2022). Some tools applied augmented reality (AR) and VR techniques to provide an immersive experience to prospective dancers (Anderson et al., 2013; Piitulainen et al., 2022). There is a recent work on the reconstruction of a professional ballet instructor's 3D choreography data acquired through motion capture in a VR ballet training environment (Choi et al., 2021). With the system, the learners can replay specific dance

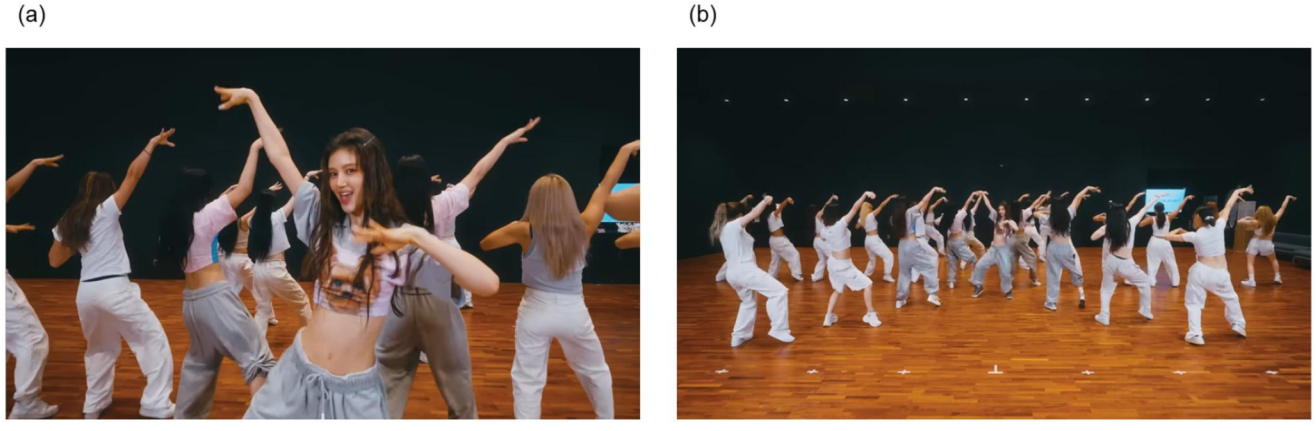

Figure 1. Characteristics of commonly used dance practice videos (2D) (a) example scene in a video (https://www.youtube.com/watch?v=wU2siJ2c5TA&ab_channel=NewJeans) where the view shifts to highlight a specific body part of a dancer; (b) Another example scene in a video (https://www.youtube.com/watch?v=qsHvdRA0a6o&ab_channel=NewJeans) where the view is fixed to show the entire body of the dancer (Note that the two videos are different videos of the same dance).

movements from any angle they want. It allows users to gain a deeper understanding of spatial orientation and body mechanics.

As mentioned before, today's dance learning method is evolving alongside these technologies of computer vision to help more people learn and enjoy dance (Cao, 2023). In this work, we aim to build on the strengths of existing research while addressing the needs identified in our preliminary research.

2.2. 3D Human reconstruction technology

With recent advances in deep learning, researchers are actively working on 3D human reconstruction (3D HR) technology that not only recognizes people in 2D images but also reconstructs them into 3D meshes (Tian et al., 2023). Many studies have used parametric human body models such as Skinned Multi-Person Linear Model (SMPL) (Loper et al., 2023) and SCAPE (Anguelov et al., 2005) to obtain 3D avatars of humans from 2D images (Cheng et al., 2018). Specifically, each parameter is needed to determine the position and orientation of major human joints and the shape of the mesh relative to the body shape. Simultaneously, the global position of the 3D avatar must be determined by predicting the intrinsic and extrinsic parameters of the camera that captures the target. If multiple targets are being tracked, 3D HR typically requires cropping an area that corresponds to each target.

Most deep learning-based 3D HR models are designed with an encoder-decoder structure. For the encoder, convolutional backbones such as ResNet (He et al., 2016) and HRNet (Wang et al., 2020) have been used in studies including HMR (Kanazawa et al., 2018), PyMAF (Zhang et al., 2021), and HUND (Zanfir et al., 2021) for feature extraction from the images. In addition, Xu et al. (2020) and Kocabas et al. (2021) have been proposed for robust parametric model estimation. As the most recent model, HMR 2.0 (4D-humans) (Goel et al., 2023) utilized ViT (Dosovitskiy et al., 2020) and a vanilla transformer (Vaswani et al., 2017) as the encoder and decoder, respectively. They performed SMPL parameter estimation and reported excellent reconstruction performance. The model takes an image containing people as input, estimates the parameters of SMPL, and recovers a 3D avatar consisting of both 6890 vertices and a 3D skeleton structure with 23 joints. Specifically, the first output of 4D-humans based on SMPL consists of $\theta \in R^{23\times3\times3}$, which represents pose information by determining the position of each joint in terms of orientation relative to its parent joint. Next, $\beta \in R^{10}$, which determines the mesh shape of the SMPL avatar. Lastly, $\pi = (R \in R^{3\times3}, t \in R^3)$, which represents the translation and rotation information of the camera that captures the target. Recent work, including 4D-humans models (Goel et al., 2023), performs not only the 3D HR but also the tracking for multiple targets in a time series. These 3D HRT models have evolved to utilize architectures including recurrent neural networks (RNN) (Doersch & Zisserman, 2019; Kocabas et al., 2020) and transformers (Rajasegaran, 2021; Shen et al., 2023) to robustly track moving targets that are occluded or disappear in temporal images. They have exploited various temporal features that utilize information including optical flow (Lee & Lee, 2021; Li et al., 2022), parameters of the parametric human body model (Kanazawa et al., 2019; Luo et al., 2020), and color information of clothing (Kanazawa et al., 2018).

One of the major challenges in 3D HR is handling occlusions, where parts of the target person are obscured by other people or objects in the scene. In the context of dance learning, occlusion is a significant problem, as dancers often move in close proximity to each other, leading to frequent occlusions in the video footage. Our preliminary survey revealed that the loss of information caused by occlusions is one of the main limitations of using 2D videos as a learning tool for dance. To address this issue, the 3D HR technology employed in our study is designed to be robust against occlusions. The 3D HR model is trained on a large dataset using a strategy that involves masking specific joints of the target in the image, simulating occlusion scenarios. This training approach enables the model to reconstruct the most plausible pose even in the presence of occlusions. By leveraging the occlusion-resilient 3D HR technology, our DanceSculpt system aims to compensate for the loss of information caused by occlusions in 2D dance videos. The reconstructed 3D avatars provide learners with a complete view of the target dancers' movements, even when parts of their bodies are occluded in the original video. This allows learners to observe and analyze the dance movements from multiple perspectives, facilitating a more comprehensive understanding of the choreography and techniques.

Despite the significant advancements in 3D HR technology, existing algorithms still face challenges in accurately estimating human body poses under certain conditions, such as occlusion, multiple individuals, motion blur, and moving cameras. These challenges can lead to inaccuracies in the reconstructed SMPL parameters, which may affect the usability of systems that rely on these algorithms. In the context of dance learning, the accuracy of the reconstructed 3D avatars is crucial for learners to observe and analyze the dance movements effectively. Inaccurate or distorted postures of the 3D avatars may hinder the learning process by providing misleading information. Therefore, it is essential to consider the potential impact of these inaccuracies on the users' learning experience. To address this concern, we have adopted the state-of-the-art 3D Human Reconstruction & Tracking (3D HRT) model called 4D-humans (Goel et al., 2023) in our DanceSculpt system. This choice was made after comparing various models in terms of their accuracy and performance (mean-per-joint-position error of 70.0 mm on the 3DPW dataset and 44.8 mm on the Human3.6M dataset). While 4D-humans represents the current best practice in 3D HR technology, it is important to note that while the 3D HR technology employed in our study is designed to handle occlusions, there are still technical limitations to overcome in extreme cases. For example, when a significant portion of the target dancer is occluded for an extended period, or when the target is completely absent from the camera view, the reconstruction may be less accurate. In the computer vision community, researchers are actively

working on addressing these limitations through advanced tracking methods that incorporate smoothing or compensation techniques in post-processing (Muhammad et al., 2022). In our study, we acknowledge these limitations and focus on demonstrating the potential of using occlusion-resilient 3D HR technology as a learning tool for dance.

As recent AI research has presented high-quality human body reconstruction, various support tools have been also developed. In the case of AIFit (Fieraru et al., 2021), they reconstructed user and trainer movements in a fitness environment in 3D and identified deviations between their movements. Since the reconstructed 3D models can be visualized, manipulated, and analyzed in ways that are not possible with 2D images, they facilitate a more comprehensive understanding of human movement and form. In the case of PoseCoach (Liu et al., 2022), they developed a system to provide feedback on running posture to amateurs. They compared the running posture of professionals and amateurs to provide quantitative feedback and improved the learners' performance. In this study, we leverage the 3D HRT technology, specifically 4D-humans (Goel et al., 2023), that generates ID-assigned and tracked multiple SMPL models from videos including multiple dancing people. It enables visualizing the 3D avatars of reference dancers to dance learners. The learners can analyze the movements of reference dances in the videos and interpret the reference dancers' complex movements. In addition, by comparing the recognized SMPL parameters of both reference and learners, we can provide performance feedback to users in terms of dance elements. By providing relevant dance information and feedback to users through the proposed DanceSculpt system, it enhances the effectiveness of their dance learning.

3. Methodology: The dancesculpt system

The DanceSculpt (DS) system is designed to support users' dance learning by providing not only 2D video frames but also detailed 3D information of dancers analyzed from reference videos. The system addresses the key challenges and requirements identified through a preliminary survey and analysis of dancers' learning processes, such as the need for detailed visual information, synchronized multi-view representations, and quantitative performance feedback. The DS system consists of two main components: the dancer viewer (Figure 2(a)) and the dance feedback tool (Figure 2(b)). These components are integrated into a comprehensive system with graphical user interfaces (GUIs) to assist dance learning in two typical stages: the dance familiarization stage and the monitoring stage.

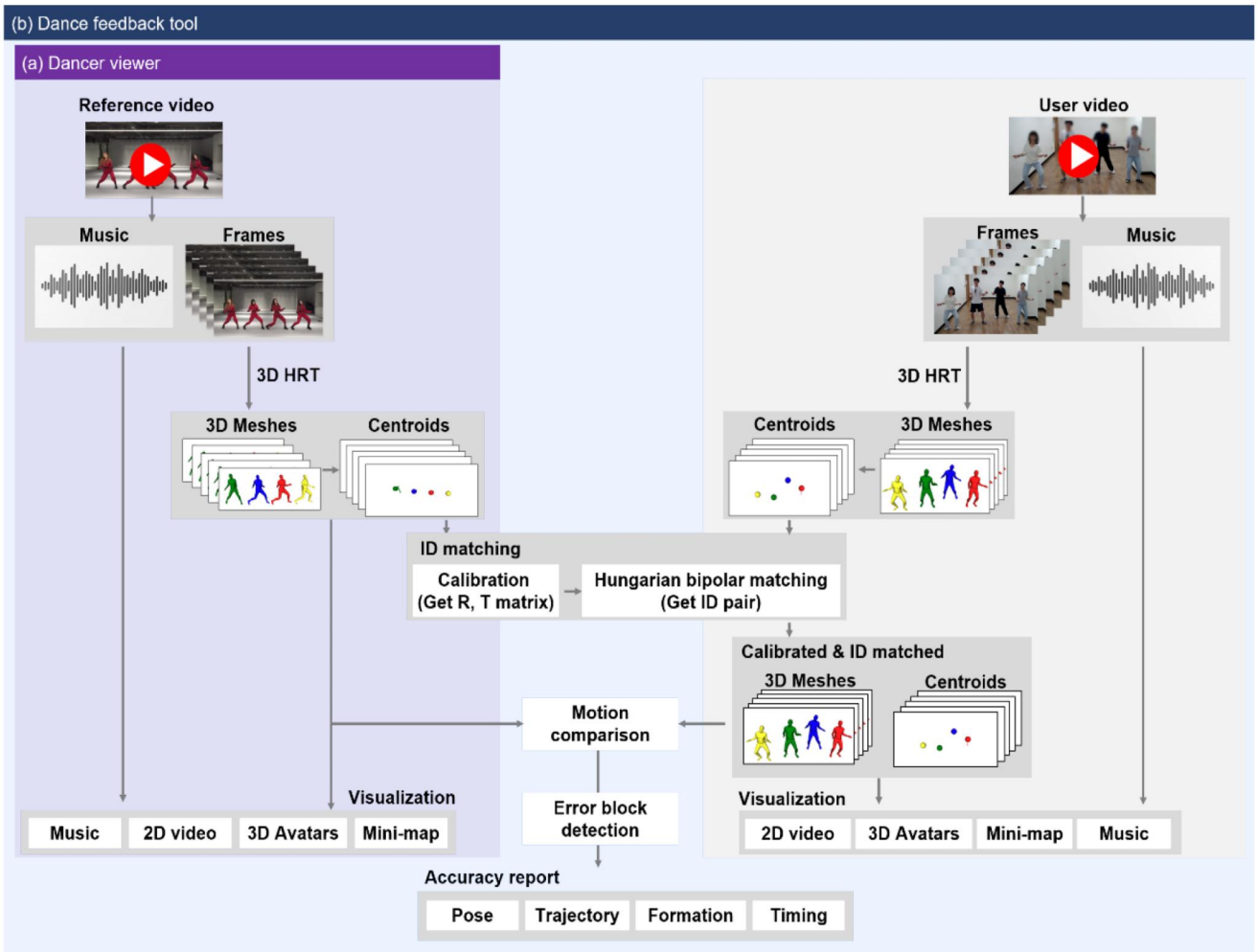

Figure 2. The overall architecture of the DanceSculpt system (a) dancer viewer; (b) dance feedback tool.

One of the key technical contributions of the DS system is the development of a robust pipeline for reconstructing 3D avatars from both the reference video and the user's dance performance. This involved calibrating the avatars to a common coordinate system, ensuring accurate alignment and synchronization. To overcome the limitations of the baseline 3D HRT method in capturing fine-grained details and ensure the visual quality of the reconstructed models, we incorporated additional post-processing techniques. Specifically, we employed Kalman filtering to smooth the 3D avatars and reduce noise, resulting in enhanced clarity and visual fidelity. Furthermore, we designed and implemented intuitive user interfaces that allow users to interact with and manipulate the 3D avatars, providing a rich and immersive learning experience. The dancer viewer enables learners to observe and analyze dance movements from multiple perspectives, while the dance feedback tool provides quantitative evaluation of learners' dance performances based on pose accuracy, trajectory, formation, and timing. We optimized the 3D reconstruction pipeline for improved efficiency and scalability, enabling the processing of longer dance sequences and multiple dancers. Additionally, we developed efficient data structures for real-time synchronization and rendering of 3D avatars, ensuring a smooth and responsive user experience. Note that all frames of 2D video, 3D avatars, and music are presented to the user in a synchronized manner, facilitating a comprehensive understanding of the dance choreography and techniques.

3.1. Dancer viewer

For each frame of the input video, the DS system uses the 3D HRT technique to segment the dancers and convert them into 3D SMPL avatars. Specifically, 3D HRT first utilizes a human detection model to extract regions of the input image corresponding to humans. In case of 4D-humans adopted in this study, the Detectron 2.0 (Wu et al., 2019) was used for human detection. Then, the 3D HRT estimates the parameters of human parametric model for every extracted image of human. The encoder of 4D-humans is ViT based structure and encoded features and learnable SMPL query tokens feed into standard transformer followed by the MLP layer. Finally, the output of 3D HRT is the parameters of SMPL model as described in Section 2.2. In addition, the recent 3D HRT model includes the function of tracking of dancers across multiple frames and assigns them ID values. There are many kinds of reasonable features for tracking, the 4D-humans utilizes the feature associated with the pose, location, and appearance of the target. It is same approach of the author's previous works, PHALP (Rajasegaran et al., 2022). They train a structure that predicts the next feature using the previous feature based on the transformer model. Then, tracking is performed by comparing predicted values and actual observed values. As mentioned in Section 2.2, considering these factors, we determined that the 3D data provided by 4D-humans was the most suitable for investigating the impact of 3D avatars on dance learning. Therefore, we were able to construct our system based on 4D-humans without requiring additional modifications to the 3D HRT component. However, it is important to note that while 4D-humans operates on a frame-by-frame basis, it does not address the reconstruction of multiple objects within a single image in the same coordinate system. Therefore, in the process of implementing the system by applying 3D HRT, we had to perform additional setting work for each video to modify the parameter required for depth estimation in order to reconstruct the avatar in the same coordinate system.

Figure 3(a) shows an example GUI of the dancer viewer. It shows the input video, a 3D panel displaying avatars extracted from the video, a mini-map to briefly present the trajectory and formation of dancers, and a standard video controller. A user can manipulate the viewpoint of the 3D panel as described in the Figure 3. The detail of each components of dancer viewer are explained below.

3.1.1. 3D Avatar

Each reconstructed 3D avatar is represented by a different color according to its ID value (e.g., red, blue, green, yellow). The user can select specific 3D avatar what they want to highlight, where the remaining 3D avatars are visualized as translucent by adjusting the alpha value to help the user focus. In addition, the user can rotate (drag), zoom in, and zoom out (wheel control) to move the camera viewpoint with simple mouse controls to observe some dance movements in detail.

3.1.2. Mini-map

The mini-map provides users with the information of trajectory and formation of multiple dancers in a video more intuitively. It visualizes the centroids of dancers as "dots" in the top view. The movement path of the dots is presented as "lines" for the last 3 s (30 frames per second (fps) * 3 seconds = 90 frames) to show their trajectory. The Mini-map feature is located in the top right corner of the dancer viewer and visualizes the relative positions of the 3D avatars. This intuitive visual effect conveys further information about positional movement and formation.

3.1.3. Video controller

The DS system provides 3D dance information for every frame of the input video, where typical playback functions of common video players are implemented with GUI widgets using the vedo library.[1] The user can control DS through keyboard and mouse control like typical 2D video player: (1) play/pause—play or pause the data with the same speed as the original video; (2) 10 seconds back or forward—move the frame 10 seconds back or forward; (3) repeat—repeat the frames set by the user (key "[', ']"); (4) playback speed control—change the speed of play (key "+" for faster, "−" for slower); (5) slide bar—move to the visualized frame selected by the user; and (6) mirror—all displays flip left and right (key "m").

Figure 3. UI examples of the DS system (a) dancer viewer; (b) dance feedback tool.

3.2. Dancer feedback tool

With the dance feedback tool of the DS system, the users can compare two corresponding dancers in the two videos (one from reference, the other from user's). When users input dance videos of both their and reference for getting some feedback, our DS system compares dancers in the two videos to measure how similarly dancers perform the same dance. In order to measure the similarity of dancers between two videos, the DS system matches the corresponding dancers assigned with the same ID in both videos, and evaluates their dance performance in terms of pose, trajectory, formation, and timing. To automate this process, we first perform a coordinate system calibration to place the user and the reference dancer in the same coordinate system using a method that applies singular value decomposition (SVD) on the object trajectory (Lee et al., 2022). We first selected the center points of each 3D avatar as the components of the trajectory pairs. Then, we calculate the rotation and translation matrix between trajectory pairs using SVD. Next, we matched the ID pairs between the user and the reference using Hungarian bipolar matching between groups of mesh centers located in the same coordinate system (Kuhn, 1955). Finally, we obtained reference meshes matched with corresponding to each user in the calibrated coordinate system. After calibration process, the tool displays 2D videos and 3D panels of both reference and user, provides the evaluation of the user's dance performance compare with the reference, as shown in Figure 3(b). Also, the user can select for highlighting specific avatars present in the 3D panel what they want to focus in the comparing at the same way with dancer viewer.

For the quantitative evaluation of the dance performance, a motion analysis algorithm is proposed to compare the dance between the user and the reference. The following is the implementation of the motion analysis algorithm. First, we match the series 3D data (i.e. SMPL parameter) of users and references along the time axis using the Dynamic Time Warping (DTW) method. We defined the cost metric for the time axis matching as the cosine similarity between two one-dimensional vectors consisting of pose data (SMPL parameters, 23 joint orientation values) from every 3D avatar of reference and user. By considering all objects while performing each frame matching, we are able to achieve a more robust matching against errors occurring in specific subjects or at specific joints. All dance analytical elements are calculated by comparing matched frames. In addition, analytical elements other than formation are calculated on a per-person basis. Next, we explain how to calculate each analytical element.

3.2.1. Pose accuracy

The pose accuracy between a user and the reference dancer is computed by measuring the cosine similarity between their poses. It is mainly adopted matric as a loss in training of 2D pose estimation model (Xiao et al., 2018). As an input, we utilize the pose information of the SMPL parameters extracted from the 3D HRT model. Specifically, we select the joints corresponding to the limbs in the similarity

calculation process (left and right shoulders, elbows, wrists, pelvis, knees, and ankles), but excluded the joints corresponding to the torso, which have relatively little variation in motion since they move as a single rigid body.

3.2.2. Trajectory accuracy

The DS system evaluates trajectory accuracy in terms of formation, considering only the orientation of the target relative to the other dancers. We compute the unit vectors for all dancers in the group from the target using the center of the mesh. We then flatten them into a one-dimensional vector to obtain the position vector for a target. Finally, we define the trajectory similarity of the user as the cosine similarity between the user-reference dancer position vectors.

3.2.3. Formation accuracy

Formation accuracy is the average value of the trajectory similarity for all users. The users can find the cause of low formation accuracy by exploring the trajectory accuracy for each user.

3.2.4. Timing

Timing is calculated for each frame by performing an additional DTW match for each user. If the matched frame number of reference is larger or smaller than the user's, the user's action is annotated as faster or slower than the reference.

Four dance analytic elements are presented as graphs as shown in Figure 3(b). To intuitively convey to the user where they are wrong, error information is displayed for each analysis element in the form of an error block and highlighted as a red area on the graphs. If the displayed frame corresponds to an error block, it encourages the user to stop playing and focus on the 3D penal or 2D image. In case of pose accuracy, we additionally provide textual feedback describing incorrect body parts (e.g., both hands and legs). This textual feedback is provided as flags associated with the corresponding 3D mesh (i.e., the avatar). The error block is defined as the region that corresponding accuracy is lower than threshold and persists more than one second (=30 frames). The thresholds for each analytic element are defined as the holding time of the error for Timing and the accuracy (in percentage) for the rest of the elements, respectively. We assumed that the users may have different levels of desire to improve detected errors for different groups. For example, the beginner can adjust the threshold to a high level to highlight and observe only relatively large errors. Therefore, we provide the controllable threshold for error detection via slide bar widget.

4. Experiments

4.1. Experimental design

We recruited 30 university students (16 females, 14 males) with a mean age of 23.2 years (18–33 years, $SD = 4.14$) and a mean dance experience of 1.7 years (1–4 years, $SD = 0.83$), consisting of eight groups (six groups of four and two groups of three). All participants had not taken any professional dance lesson and used to watch dance videos to learn to dance. For the experiment, we selected two K-pop female group dance songs (Song A: Aespa—Savage,[2] Song B: Aespa—Spicy[3]) that the participants had never practiced. These songs include a variety of pose, trajectory, and formation changes. Participants were asked to learn to dance that corresponded to each song.

Each group participated in two sessions on different days (Figure 4). In the first session, they learned and practiced the two songs (Stage a). The second session was composed of monitoring (Stage b) and another practice (Stage c). In this paper, we utilized a counterbalanced manner to balance the experiment across subjects and conditions in order to address the inter-subject variability (e.g., learning effect) resulting from individual differences in learning patterns. Each groups practiced Songs A and B through both CP and DS in a counterbalanced manner (i.e., some group learned song A with DS and the other with CP). For CP, they used the VLC media player,[4] an open-source library, as the conventional video player. The participating groups performed Stages a, b and c for 40 min each (20 min per song). They

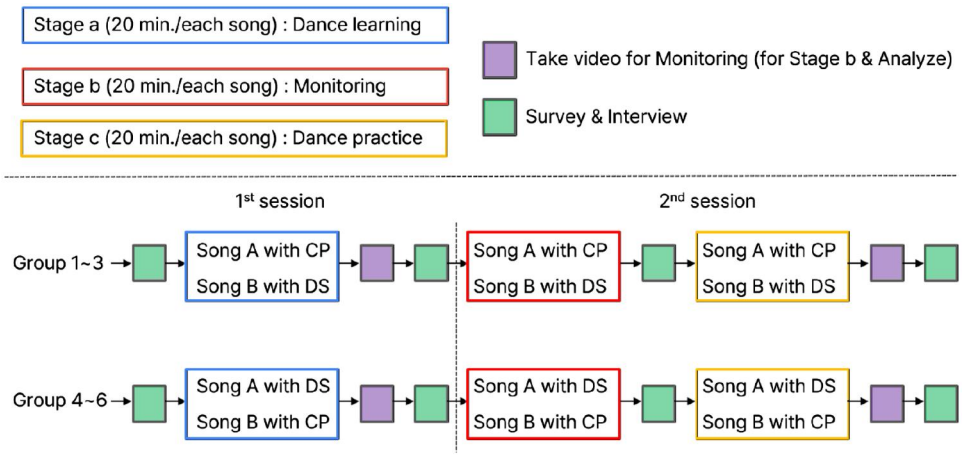

Figure 4. A summary of the experimental process over participants.

were asked to submit final dance videos of both Song A and Song B after Stages a and c (Sa/A, Sa/B, Sc/A, Sc/B), respectively. In total, 24 videos were submitted from the six groups (=6 groups × 4 videos). The study was conducted in accordance with the Declaration of Helsinki, and approved by the institutional review Board(IRB) of Gwangju Institute of Science and Technology, Republic of Korea (protocol code 20210806-HR-62-02-02 and date of approval 17 August 2021).

4.2. Data collection and analysis

The participants conducted their sessions in a prepared dance practice space. We recorded the entire process with both front and side views using common webcams (1080p, 30fps) to observe their behaviors. They were asked to fill out questionnaires and participate in interviews. We analyzed behavioral responses and measured the performance improvement when using the DS system and the CP.

4.2.1. Dance performance

To evaluate participants' learning performance, we analyzed submitted dance videos (Sa/A, Sa/B, Sc/A, Sc/B). The dance performance was then evaluated by using the motion comparison algorithm of the DS system. Statistical analysis was performed using the Wilcoxon signed-rank test for the measured items.

4.2.2. User experience

To analyze the user experience, we developed a questionnaire inspired by the System Usability Survey (SUS) (Bangor et al., 2008) and the User Experience Questionnaire (UEQ) (Laugwitz et al., 2008). This questionnaire was answered using a 5-point Likert scale. The detailed questionnaire items are included in Figure 6.

4.2.3. Behavioral responses

We transcribed the recorded conversations of the participants with time stamps. We keyword-tagged the collected text data using ATLAS.ti,[5] a qualitative research software tool developed by Scientific Software Development GmbH. Finally, we measured the frequency of behavioral responses using the categories shown in Table 1, and the results are described in detail in Table 2.

5. Results

Figure 5 shows the change in performance of every participant from Stage a (Sa = {Sa/A, Sa/B}) to Stage c (Sc = {Sc/A, Sc/B} by measuring the temporal and spatial deviation of dance between the all users and the corresponding reference dancers.

To evaluate the participants' initial learning rate, we first investigated their performance in Sa. With DS, the performance of all factors in Sa was higher compared to CP by an average of 7.37%. Specifically, the difference was 10.50% and 10.61% for Trajectory and Formation. This indicates that the 3D information provided in DS allowed users to quickly recognize the information regarding Trajectory and Formation compared to CP. For Pose, the performance was improved by an average of 3.41% with DS, although not significant.

Comparing the performance improvement between Sa and Sc, CP improved by 4.77% compared to 6.81% for DS when looking at all factors except Timing. Given that DS

Table 1. Behavioral response variables and signals.

Behavioral variables		Signals
Mastery of dance elements (Pose/Trajectory/Formation/Timing)	Question	Verbal and physical interaction through questions (e.g., "Which way should I turn my hand?"). This typically occurs when users are not familiar with the dance elements.
	Confirmation	Physical interactions such as verbal responses and demonstrations when users have mastered the dance elements correctly (e.g., "I should go forward like this, and you should go backward").
Emotion	Discontent	Expressions of difficulty or annoyance (e.g., sighing, grumbling, showing frustration).
	Confidence	Expressions of self-assurance or mastery (e.g., laughter, applause).
	Passion	Statements of motivation to learn more (e.g., "Let's take a closer look").

Table 2. The average frequency of behavioral responses observed in each stage (standard deviations in parentheses).

Behavioral response variables			Stage a		Stage b		Stage c		Total avg.	
			CP	DS	CP	DS	CP	DS	CP	DS
Mastery of dance elements	Question	Pose	37.25 (18.41)	47.13 (19.37)	24.75 (11.18)	30.00 (22.08)	12.88 (5.99)	16.50 (4.24)	24.96 (15.98)	31.21 (20.78)
		Trajectory	6.25 (2.55)	11.88 (7.10)	16.38 (6.63)	12.00 (3.02)	20.00 (10.35)	22.88 (11.74)	14.21 (9.13)	15.58 (9.37)
		Formation	4.25 (2.82)	11.50 (6.70)	20.13 (5.57)	20.13 (9.46)	17.00 (9.25)	15.50 (7.75)	13.79 (9.15)	15.71 (8.49)
		Timing	2.75 (1.75)	2.25 (1.16)	1.50 (1.31)	11.63 (6.32)	17.75 (9.25)	19.38 (4.53)	7.33 (9.19)	11.08 (8.37)
		Sum of Question	50.50	72.75	62.75	73.75	67.63	74.25	60.29	73.58
	Confirmation	Pose	11.88 (8.89)	19.13 (8.08)	11.50 (6.52)	14.88 (3.18)	10.00 (2.51)	5.75 (4.20)	11.13 (6.29)	13.25 (7.80)
		Trajectory	6.50 (5.21)	18.38 (8.86)	4.50 (2.67)	18.50 (7.62)	3.13 (1.96)	15.38 (5.18)	4.71 (3.69)	17.42 (7.20)
		Formation	4.75 (2.38)	12.00 (6.41)	5.25 (4.65)	17.25 (8.43)	3.88 (2.30)	15.00 (7.17)	4.63 (3.20)	14.75 (7.39)
		Timing	0.88 (0.35)	2.13 (0.83)	1.50 (1.07)	3.13 (1.64)	3.63 (2.50)	4.00 (1.85)	2.00 (1.93)	3.08 (1.64)
		Sum of Confirmation	24.00	51.63	22.75	53.75	20.63	40.13	22.46	48.50
Emotion		Discontent	34.50 (26.12)	34.63 (23.93)	6.88 (5.57)	13.88 (5.82)	11.50 (4.90)	4.38 (1.60)	17.63 (19.41)	17.63 (18.76)
		Satisfaction	12.00 (10.61)	5.50 (4.41)	7.50 (5.50)	11.75 (14.68)	5.75 (1.91)	11.13 (4.36)	8.42 (7.20)	9.46 (9.25)
		Passion	2.88 (1.81)	6.75 (2.55)	6.00 (3.51)	10.38 (7.37)	1.25 (0.89)	5.13 (2.59)	3.38 (3.00)	7.42 (5.06)

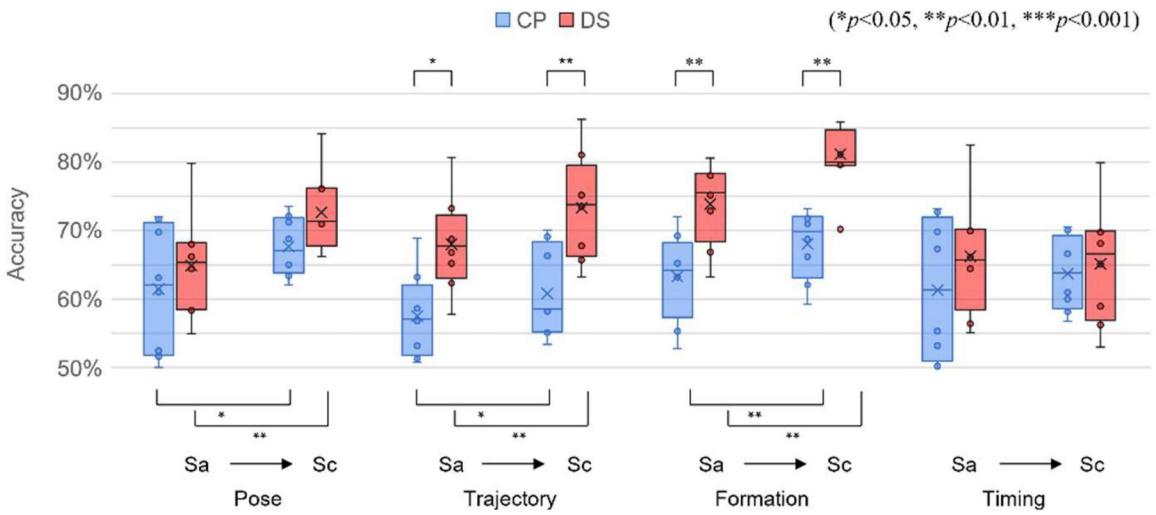

Figure 5. Comparison of participants' dance performance by learning methods.

improved by more than 2.04% on average compared to CP suggests that reviewing with the dance feedback tool is consistently effective in the learning process. The performance change of the Pose from Sa to Sc was 1.57% higher for DS (7.80%) than CP (6.23%). For the Trajectory, DS (5.33%) was 2.02% better than CP (3.30%) on average. For the Formation, DS (7.30%) was more effective than CP (4.77%). In the case of Timing, there was no significant difference (DS was 3.57% lower than CP).

We statistically analyzed the questionnaire answers to ensure that the results of the performance comparison matched the participant's self-reported experience. As shown in Figure 6, the participants found that DS allowed them to observe and understand the new dance without missing key details (B. Effectiveness, Avg. = 4.27), and helped them master the new dance (C. Usefulness, Avg. = 4.17). Confidence in dancing increased with both CP and DS after the training, but especially with DS, with confidence rising to Avg. = 3.67. While not a statistically significant difference compared to CP (Avg. = 3.30), this analysis confirmed the possibility that providing sufficient information about dancing can positively contribute to the confidence of participants learning to dance.

We further asked specific questions to determine how effective and useful participants found DS to be for each dance element (Pose, Trajectory, Formation, and Timing), as shown in Figure 6(D&E). Regarding whether they were able to observe all necessary dance information (D. Effectiveness), participants answered that they were able to recognize the key details of Pose, Trajectory, and Formation compared to CP. Regarding whether it helps them to learn the dance (E. Usefulness), they felt DS is superior to CP in all the dance elements and we found statistically significant difference in Trajectory and Formation. *P3* stated: "*I think it was especially good when I checked the lineup. It was nice to be able to easily compare the trajectory using the mini-map, and I think it was something difficult to catch in conventional dance videos.*" In addition, *P9* stated: "*It was helpful to be able to check in 3D which way I needed to turn.*" However, there were no significant differences in Pose and Timing, which is consistent with the trend of the participants' performance comparison results (Figure 5). This result indicates that further system improvements and follow-up studies may be needed depending on the type and difficulty of the target dance.

As shown in Figure 6(F-1), participants learned to use the DS system without much difficulty, compared with a typical video player. A few participants wanted a more intuitive and simple UI design for 3D display and interaction, e.g., *P22* stated: "*I'm not used to seeing my own 3D avatar in three dimensions while rotating it.*" Despite the need for participants to learn and familiarize themselves with it, participants rated DS positively in terms of functionality and reusability. Additionally, we further explored satisfaction with specific DS functions, where they were generally satisfied with the information provided by the 3D panel, the mini-map feature, and the dance feedback tool (Satisfaction | Avg. = 4.03, $SD = 0.61$/Avg. = 3.97, $SD = 0.32$/Avg. 4.03, $SD = 0.67$).

To further validate the usefulness of DS, we analyzed participants' behavioral responses (Table 2). Across all stages, the amount of Question was not significantly different, averaging 13.29 more in DS than CP, but the amount of Confirmation was 26.04 more in DS than CP. In particular, the amount of Confirmation for Formation and Trajectory was higher by 12.71 and 10.13, respectively. However, no significant differences were observed in the mean values for Pose and Timing. This is consistent with the results of the quantitative evaluation, given the definition of Confirmation. Unusually, participants' Satisfaction with DS was lower in Sa than in CP by 6.50. This is due to the relatively large amount of information that DS conveys to the users, which puts a strain on cognition. In fact, *P22* said, "*It was hard to realize that there are more elements to consider than I thought.*" However, as confirmable in Sc of Table 2, users were observed 5.38 more Satisfaction with DS than CP. Synthetically, DS was 1.04 more satisfied on average than CP in Satisfaction.

To summarize the results, DS can provide users with high-quality information about dance elements, resulting in higher overall satisfaction and consequently better dance performance for users compared to CP. *P13* stated: "*At first I thought it was just silly, but as time went on, I could see*

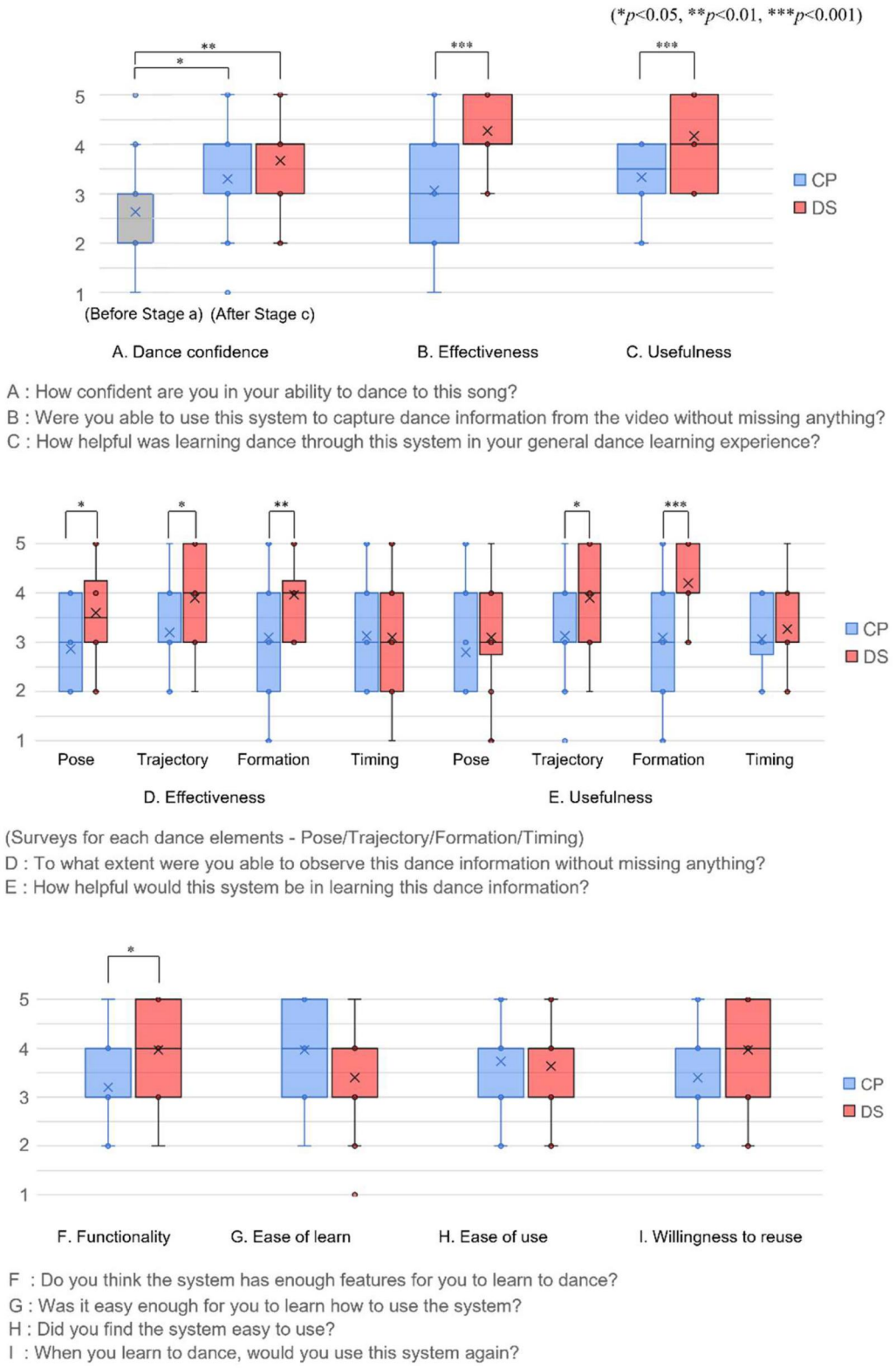

Figure 6. Comparison of user experience of dance learning approaches.

myself comparing the 2D and 3D screens and getting feedback on my dancing and posture."

6. Limitations and future work

Although our DanceSculpt system demonstrates the potential of using occlusion-resilient 3D HR technology for dance learning, there are several limitations and considerations for future work. Firstly, the current 3D HR technology employed in our system does not support real-time processing due to the computational complexity of the AI models. To improve the usability of the system in practical applications, it is necessary to optimize the 3D HR technology for faster processing. Further research on model compression

and lightweight architectures may enable more efficient and real-time operation of the DanceSculpt system. Secondly, while the current 3D HR technology can effectively reconstruct large movements and general poses based on the SMPL parameters, it still lacks the ability to capture fine-grained details such as finger movements and facial expressions. In our user study, participants mentioned that the large gesture representation provided by the DanceSculpt system was sufficient for learning the overall dance choreography. However, for professional dancers or those learning more intricate dance styles, the absence of fine-grained representations may limit the usefulness of the system. To address this limitation, future work could explore the integration of additional computer vision techniques, such as specialized models for reconstructing detailed facial expressions or precise estimation of 3D hand poses. By combining these advanced recognition models with the existing 3D HR technology, the DanceSculpt system could provide even more comprehensive and precise 3D dance information. Furthermore, as discussed earlier, there are still technical challenges in handling extreme occlusion cases, such as when a significant portion of the target dancer is obscured for an extended period or when the target is completely absent from the camera view. Future research could investigate advanced tracking and post-processing techniques, such as smoothing or compensation methods, to improve the reconstruction accuracy in heavily occluded situations.

Another important consideration is the potential impact of inaccuracies in the 3D human reconstruction algorithm on the usability of the DanceSculpt system. As mentioned earlier, the accuracy of the reconstructed SMPL parameters is crucial for learners to effectively observe and analyze dance movements. During our user study, participants reported that minor parameter errors did not significantly hinder their learning process, as they were aware of the physical differences between themselves and the reference dancer. However, they also mentioned that severely distorted postures in the 3D avatar could interfere with their observation and understanding of the dance movements. To mitigate this issue, we provided users with both the 3D avatar viewer and the original 2D video, allowing them to refer to the video when encountering distorted postures in the 3D avatar. Nevertheless, we acknowledge that the impact of distortions may be more significant for highly complex dance compositions, and this limitation should be considered when applying the DanceSculpt system in various dance learning scenarios. Future research could investigate the use of advanced 3D HRT models, such as SLAHMR (Ye et al., 2023) and TRACE (Sun et al., 2023), which have shown promise in improving accuracy in environments with moving cameras and motion blur. Incorporating these models into the DanceSculpt system could enhance its robustness and usability, particularly in challenging dance performance settings. Furthermore, future work could explore the development of error detection and correction mechanisms to identify and mitigate the impact of inaccuracies in the reconstructed 3D avatars. This could involve the use of machine learning techniques to automatically detect and filter out distorted postures, or the implementation of user feedback systems to allow learners to report and correct errors in the 3D avatars.

In addition to these technical enhancements, future work could also explore the application of the DanceSculpt system in various dance education settings, from beginner-level classes to professional training. Conducting user studies with diverse groups of dancers and instructors would provide valuable insights into the system's effectiveness and usability across different skill levels and dance styles. This feedback could inform further refinements and customization of the system to meet the specific needs of each target user group. We will then consider the additional adaptation of a number of useful evaluation metrics used in many recent dance-focused studies (Li et al., 2021, 2023; Zhou et al., 2023).

In conclusion, while the DanceSculpt system demonstrates the potential of using 3D HR technology for dance learning, it is important to recognize the limitations posed by the accuracy of the underlying algorithms. By acknowledging these limitations, providing users with complementary resources, and continuously improving the system based on the latest advancements in 3D HR technology, we can work towards creating a more robust and reliable tool for dance education and analysis. Overall, while the DanceSculpt system presents a promising approach for dance learning using occlusion-resilient 3D HR technology, there are opportunities for future research and development to address current limitations and expand its capabilities (Hanna, 2008).

7. Conclusion

On the recruited eight groups of dance learners (a total of 30 students), we have validated the proposed DS system and found its potential benefits to their dance learning. Compared to practicing dance based on typical videos, the 3D information reduced errors in posture and movement by improving the overall the quality of dance practice and their dance learning experience. Also, the feedback we collected during the experiment in this study was positive in general. The users found the DS system is easy to navigate, while the controllable viewing angle gave them a more comprehensive understanding of the dance routine. They also found the dance feedback tool particularly useful, as it allowed them to identify and correct mistakes. The improvement in dance skills and positive user feedback demonstrate our system's potential as a powerful dance learning aid. In addition to the results, the study also reveals topics for further research. One potential direction is to investigate the effectiveness of the DS system in educational environment where different styles of professional dance are taught, from ballet to hip-hop or modern dance. We also expect to find insights into how the technology might interact with novice dancers as well as experienced professionals.

Notes

1. https://vedo.embl.es/.
2. https://youtu.be/jVkHUbF_rfg?si=sRJw30etN8KE6h8a.
3. https://youtu.be/pPoYQPcBUQ4?si=IECOFuikNyCZMXj-.
4. https://www.videolan.org/vlc/index.ko.html.
5. https://atlasti.com/.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the 'Project for Science and Technology Opens the Future of the Region' program through the INNOPOLIS FOUNDATION funded by Ministry of Science and ICT (Project Number: 2022-DD-UP-0312); Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01842, Artificial Intelligence Graduate School Program (GIST)).

ORCID

Sanghyub Lee http://orcid.org/0000-0001-8363-3054
Woojin Kang http://orcid.org/0000-0001-9765-7479
Duk-Jo Kong http://orcid.org/0000-0001-8674-8981

References

Anderson, F., Grossman, T., Matejka, J., & Fitzmaurice, G. (2013). YouMove: Enhancing movement training with an augmented reality mirror [Paper presentation]. Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (pp. 311–320). https://doi.org/10.1145/2501988.2502045

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., & Davis, J. (2005). *Scape: Shape completion and animation of people* [Paper presenation]. ACM SIGGRAPH 2005 Papers (pp. 408–416). https://doi.org/10.1145/1073204.1073207

Aristidou, A., Stavrakis, E., Charalambous, P., & Chrysanthou, Y. (2015). Stephania Loizidou Himona. 2015. Folk dance evaluation using laban movement analysis. *Journal on Computing and Cultural Heritage*, *8*, 1–19. https://doi.org/10.1145/2755566

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Cao, X. (2023). Case study of China's compulsory education system: AI apps and extracurricular dance learning. *International Journal of Human–Computer Interaction*. Advance online publication. https://doi.org/10.1080/10447318.2023.2188539

Cheng, Z.-Q., Chen, Y., Martin, R. R., Wu, T., & Song, Z. (2018). Parametric modeling of 3D human body shape—A survey. *Computers & Graphics*, *71*(2018), 88–100. https://doi.org/10.1016/j.cag.2017.11.008

Choi, J., Massey, K., Hwaryoung Seo, J., & Kicklighter, C. (2021). Balletic VR: Integrating art, science, and technology for dance science education [Paper presentation]. 10th international conference on digital and interactive arts, Aveiro, Portuga, (pp. 1–6). https://doi.org/10.1145/3483529.3483704

Doersch, C., & Zisserman, A. (2019). Sim2real transfer learning for 3d human pose estimation: Motion to the rescue. *Advances in Neural Information Processing Systems*, *32*(2019), 12949–12961. https://dl.acm.org/doi/10.5555/3454287.3455447

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale* [Paper presentation]. The International Conference on Learning Representations (ICLR) 2021. https://doi.org/10.48550/arXiv.2010.11929

Ettina Laugwitz, T. Held., & M., Schrepp. (2008). Construction and evaluation of a user experience questionnaire. In *Proceedings HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008* (pp. 63–76). Springer. https://doi.org/10.1007/978-3-540-89350-9_6

Fieraru, M., Zanfir, M., Pirlea, S. C., Olaru, V., & Sminchisescu, C. (2021). Aifit: Automatic 3d human-interpretable feedback models for fitness training [Paper presentation]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, 2021) (pp. 9919–9928). IEEE/CVF. https://doi.org/10.1109/CVPR46437.2021.00979

Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J. (2023). *Humans in 4D: Reconstructing and tracking humans with transformers* [Paper presentation]. ICCV (International Conference on Computer Vision) 2023 (pp. 14783–14794). IEEE/CVF. https://doi.org/10.1109/ICCV51070.2023.01358

Guest, A. H. (1990). Dance notation. *Perspecta*, *26*, 203–214. https://doi.org/10.2307/1567163

Hanna, J. L. (2008). A nonverbal language for imagining and learning: Dance education in K–12 curriculum. J. *Educational Researcher*, *37*(8), 491–506. https://doi.org/10.3102/0013189X08326032

He, K., Zhang, X., Ren, S., Sun, J. (2016). *Deep residual learning for image recognition* [Paper presentation]. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).

Hong, J. C., Chen, M. L., & Hong Ye, J, the Department of Industrial Education, National Taiwan Normal University, Taiwan (2020). Acceptance of YouTube applied to dance learning. *International Journal of Information and Education Technology*, *10*(1), 7–13. 2020 https://doi.org/10.18178/ijiet.2020.10.1.1331

Hou, Y. (2022). The collision of digital tools and dance education during the period of COVID-19. In *2022 2nd International Conference on Modern Educational Technology and Social Sciences (ICMETSS 2022)* (pp. 844–852). Atlantis Press. https://doi.org/10.2991/978-2-494069-45-9_102

Kanazawa, A., Michael, J. B., Jacobs, D. W., Malik, J. (2018). *End-to-end recovery of human shape and pose* [Paper presentation]. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7122–7131).

Kanazawa, A., Zhang, J. Y., Felsen, P., & Malik, J. (2019). Learning 3d human dynamics from video [Paper presentation]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (CVPR 2019) (pp. 5614–5623). https://doi.org/10.1109/CVPR.2019.00576

Kocabas, M., Athanasiou, N., & Black, M. J. (2020). Vibe: Video inference for human body pose and shape estimation [Paper presentation]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, 2020) (pp. 5253–5263). https://doi.org/10.1109/CVPR42600.2020.00530

Kocabas, M., Huang, C.-H. P., Hilliges, O., Black, M. J. (2021). PARE: Part attention regressor for 3D human body estimation [Paper presentation]. Proceedings of the IEEE/CVF international conference on computer vision (pp. 11127–11137).

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*. *2*, 83–97. https://doi.org/10.1002/nav.3800020109

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20–21, 2008. Proceedings 4* (pp. 63–76). Springer. https://doi.org/10.1007/978-3-540-89350-9_6

Lee, G.-H., & Lee, S.-W. (2021). Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics [Paper presentation]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV, 2021) (pp. 12375–12384). https://doi.org/10.1109/ICCV48922.2021.01215

Lee, S., & Lee, K. (2023). CheerUp: A *real-time ambient visualization of cheerleading pose similarity* [Paper presentation]. Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (pp. 72–74). Association for Computing Machinery (ACM). https://doi.org/10.1145/3581754.3584135

Lee, S-h., Lee, D.-W., Jun, K., Lee, W., & Kim, M. S. (2022). Markerless 3d skeleton tracking algorithm by merging multiple inaccurate skeleton data from multiple rgb-d sensors. *Sensors (Basel, Switzerland)*, *22*(9), 3155. https://doi.org/10.3390/s22093155

Li, R., Yang, S., Ross, D. A., Kanazawa, A. (2021). *Ai choreographer: Music conditioned 3d dance generation with aist++* [Paper presentation]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021) (pp. 13401–13412). https://arxiv.org/abs/2101.08779

Li, R., Zhao, J., Zhang, Y., Su, M., Ren, Z., Zhang, H., Li, X. (2023). FineDance: A fine-grained choreography dataset for 3D full body dance generation [Paper presentation]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023) (pp. 10234–10243). https://doi.org/10.1109/ICCV51070.2023.00939

Li, Z., Xu, B., Huang, H., Lu, C., & Guo, Y. (2022). Deep two-stream video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2022)* (pp. 430–439). IEEE Computer Society. https://doi.org/10.1109/WACV51458.2022.00071

Liu, J., Saquib, N., Zhu-Tian, C., Kazi, R. H., Wei, L.-Y., Fu, H., & Tai, C.-L. (2022). *PoseCoach: A customizable analysis and visualization system for video-based running coaching* [Paper presentation]. IEEE Transactions on Visualization and Computer Graphics (pp. 1–15). https://doi.org/10.1109/tvcg.2022.3230855

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2023). SMPL: A skinned multi-person linear model. *In Seminal Graphics Papers: Pushing the Boundaries*, *2*(6), 851–866. https://doi.org/10.1145/2816795.2818013

Luo, Z., Golestaneh, S. A., Kitani, K. M. (2020). *3d human motion estimation via motion compression and refinement* [Paper presentation]. Proceedings of the Asian Conference on Computer Vision (ACCV). Springer. https://doi.org/10.1007/978-3-030-69541-5_20

Muhammad, Z.-U.-D., Huang, Z., & Khan, R. (2022). A review of 3D human body pose estimation and mesh recovery. *Digital Signal Processing*, *128*(2022), 103628. https://doi.org/10.1016/j.dsp.2022.103628

Ng, L. H. X., Tan, J. Y. H., Tan, D. J. H., & Lee, R. K.-W. (2021). *Will you dance to the challenge? predicting user participation of TikTok challenges* [Paper presentation]. Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 356–360). Association for Computing Machinery (ACM). https://doi.org/10.1145/3487351.3488276

Piitulainen, R., Hämäläinen, P., & Mekler, E. D. (2022). Vibing together: Dance experiences in social virtual reality [Paper presentation]. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1–18). https://doi.org/10.1145/3491102.3501828

Rajasegaran, J., Georgios, P., Angjoo, K., Jitendra, M. (2021). *Tracking people with 3D representations*. *Advances in Neural Information Processing Systems (NeurIPS 2021), 34*, 23703–23713. https://doi.org/10.48550/arXiv.2111.07868

Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J. (2022). *Tracking people by predicting 3D appearance, location and pose* [Paper presenation]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2740–2749). IEEE/CVF.

Shen, X., Yang, Z., Wang, X., Ma, J., Zhou, C., & Yang, Y. (2023). *Global-to-Local Modeling for Video-based 3D Human Pose and Shape Estimation* [Paper presentation]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023) (pp. 8887–8896). IEEE/CVF. https://doi.org/10.1109/CVPR52729.2023.00858

Singh, B., Beatty, J. C., & Ryman, R. (1983). *A graphics editor for benesh movement notation* [Paper presentation]. Proceedings of the 10th annual conference on Computer Graphics and Interactive Techniques (pp. 51–62). IEEE/CVF. https://doi.org/10.1145/800059.801132

Sukel, K. E., Catrambone, R., Essa, I., & Brostow, G. (2003). Presenting movement in a computer-based dance tutor. *International Journal of Human-Computer Interaction*, *15*(3), 433–452. https://doi.org/10.1207/S15327590IJHC1503_08

Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M. J. (2023). *TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments*. [Paper presentation]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023) (pp. 8856–8866). IEEE/CVF. https://doi.org/10.48550/arXiv.2306.02850

Tian, Y., Zhang, H., Liu, Y., & Wang, L. (2023). Recovering 3d human mesh from monocular images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(12), 15406–15425. https://doi.org/10.1109/TPAMI.2023.3298850

Tsuchida, S., Mao, H., Okamoto, H., Suzuki, Y., Kanada, R., Hori, T., Terada, T., & Tsukamoto, M. (2022). Dance practice system that shows what you would look like if you could master the dance [Paper presentation]. Proceedings of the 8th international conference on movement and computing (pp. 1–8). Association for Computing Machinery (ACM). https://doi.org/10.1145/3537972.3537991

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*(2017), 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(10), 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686

Whittier, C. (2006). Laban movement analysis approach to classical ballet pedagogy. *Journal of Dance Education*, *6*(4), 124–132. https://doi.org/10.1080/15290824.2006.10387325

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). *Detectron2*. https://github.com/facebookresearch/detectron2

Xiao, B., Wu, H., Wei, Y. (2018). *Simple baselines for human pose estimation and tracking* [Paper presenation]. Proceedings of the European Conference on Computer Vision (ECCV 2018). (pp. 466–481). Springer Science and Business Media. https://doi.org/10.1007/978-3-030-01231-1_29

Xu, X., Chen, H., Moreno-Noguer, F., László., & Jeni, A., Fernando De la Torre (2020). 3d human shape and pose from a single low-resolution image with self-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part IX 16*. Springer. https://doi.org/10.1007/978-3-030-58545-7_17

Ye, V., Pavlakos, G., Malik, J., & Kanazawa, A. (2023). Decoupling human and camera motion from videos in the wild. [Paper presentation]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023) (pp. 21222–21232). IEEE/CVF. https://doi.org/10.1109/CVPR52729.2023.02033

Yulu H. (2022). The collision of digital tools and dance education during the period of COVID-19. In 2022 2nd International Conference on Modern Educational Technology and Social Sciences (ICMETSS 2022) (pp. 844–852). Atlantis Press. https://doi.org/10.2991/978-2-494069-45-9_102

Z., Li, Bo Xu, H., Huang, C., Lu., & Y., Guo. (2022). *Deep two-stream video inference for human body pose and shape estimation* [Paper presentation]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2022) (pp. 430–439). https://doi.org/10.1109/WACV51458.2022.00071

Zanfir, A., Bazavan, E. G., Zanfir, M., William, T. F., Sukthankar, R., Sminchisescu, C. (2021). *Neural descent for visual 3d human pose and shape* [Paper presentation]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14484–14493).

Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z. (2021). *Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop* [Paper presentation]. Proceedings of the IEEE/CVF international conference on computer vision (pp. 11446–11456).

Zhou, Q., Cheng Chua, C., Knibbe, J., Goncalves, J., & Velloso, E. (2021). Dance and choreography in HCI: A two-decade retrospective [Paper presentation]. Proceedings of the 2021-14 CHI Conference on Human Factors in Computing Systems. (pp. 1–14). Association for Computing Machinery (ACM). https://doi.org/10.1145/3411764.3445804

Zhou, Q., Li, M., Zeng, Q., Aristidou, A., Zhang, X., Chen, L., & Tu, C. (2023). Let's all dance: Enhancing amateur dance motions. *Computational Visual Media*, *9*(3), 531–550. https://doi.org/10.1007/s41095-022-0292-6

About the authors

Sanghyub Lee received the BS degree in biomedical engineering from University of Ulsan and the MS degree in intelligent robotics from GIST in 2017 and 2019, respectively. He is currently pursuing the PhD degree with GIST. His current research interests include image processing, healthcare robotics, and pattern recognition.

Woojin Kang received the BE degree in bioelectronics from Yonsei University, in 2014, the MS degree in robotics from the DGIST, in 2016, and the PhD degree in artificial intelligence from GIST, in 2024. He is currently a Post Doc. with the School of Integrated Technology, GIST.

Jin-Hyuk Hong received the BS, MS, and PhD degrees in computer science from Yonsei University, Seoul, Korea. He is currently an associate professor with the School of Integrated Technology and the AI Graduate School, GIST. His research interests include human-computer interaction and artificial intelligence.

Duk-Jo Kong earned a BS in electronic engineering from Chungnam National University (2010), and MS and PhD in electrical engineering and computer science from GIST (2012, 2016). He was a Senior Research Scientist at GIST (2016-2021) and has been a Principal Research Scientist and Adjunct Professor since 2021.