Dance Practice System that Shows What You Would Look Like if You Could Master the Dance

SHUHEI TSUCHIDA, Kobe University, Japan
HAOMIN MAO, Kobe University, Japan
HIDEAKI OKAMOTO, Kobe University, Japan
YUMA SUZUKI, Softbank Corp., Japan
RINTARO KANEDA, Softbank Corp., Japan
TAKAYUKI HORI, Softbank Corp., Japan
TSUTOMU TERADA , Kobe University, Japan
MASAHIKO TSUKAMOTO, Kobe University, Japan

This study proposes a dance practice system allowing users to learn dancing by watching videos in which they have mastered the movements of a professional dancer. Video self-modeling, which encourages learners to improve their behavior by watching videos of exemplary behavior by themselves, effectively teaches movement skills. However, creating an ideal dance movement video is time-consuming and tedious for learners. To solve this problem, we utilize a video generation technique based on *deepfake* to automatically generate a video of the learners dancing the same movement as the dancer in the reference video. We conducted a user study with 20 participants to verify whether the deepfake video effectively teaches dance movements. The results showed no significant difference between the groups learning with the original and deepfake videos. In addition, the group using the deepfake video had significantly lower self-efficacy. Based on these experimental results, we discussed the design implications of the system using the deepfake video to support learning dance movements.

CCS Concepts: • **Human-centered computing** → **Interaction design**; *Interaction design process and methods.*

Additional Key Words and Phrases: video self-modeling, dance movements, skill acquisition, learning, deepfake

ACM Reference Format:
Shuhei Tsuchida, Haomin Mao, Hideaki Okamoto, Yuma Suzuki, Rintaro Kaneda, Takayuki Hori, Tsutomu Terada, and Masahiko Tsukamoto. 2018. Dance Practice System that Shows What You Would Look Like if You Could Master the Dance. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In physical activities, such as dance, martial arts, and sports, learners need to learn new movement skills efficiently. The skill acquisition is tedious, involving repetitive trials and errors [39]. Learning to dance is time-consuming because it involves relatively complex full-body movements with the rhythm and melody of the music. Several learning support methods and tools based on various modalities such as vision, tactile, and auditory cues have been proposed to learn dance movements efficiently [15, 22, 30]. Particularly, researchers have proposed a learning support method through

Manuscript submitted to ACM

video viewing [5]. Fujimoto et al. [12] have proposed a practice support system that presents a video of learners dancing in relatively simple postures. This system is based on the hypothesis that learning can be accelerated through a video of themselves imitating the same movements as an expert in the reference video.

A method that encourages learners to improve their movement by watching themselves performing the ideal movement is called *video self-modeling* [7]. Diane et al. [27] have confirmed improvements in trampoline maneuvers by presenting videos made by video editing software, which offered a higher skill level than learners' current ability. This result suggests that watching self-performing videos of an expert movement teaches movement skills effectively.

However, creating such videos is time-consuming and requires advanced editing skills. In addition, when dancing involves complex movements, it is difficult for learners to edit videos of themselves dancing the ideal moves to musical pieces. Fujimoto et al.'s method also has limitations in improving the final reference dance video quality in terms of the generation method. Here, we used the deep learning-based video generation technique proposed by Chan et al. [4] to automatically generate a high-definition video of learners dancing, which applies to video self-modeling. The generated video could accelerate the learning of complex moves, such as dancing.

This study proposes a learning support system that allows learning dance movements by watching a *deepfake* video of themselves mastering an expert dancer's movements. We verify whether the generated deepfake video effectively teaches the dance movements. Specifically, we construct a model by referring to Chan et al.'s method [4]. The system uses a video of learners practicing a dance for 3 min as input, and then generates a deepfake video of the learner performing the same movements as the reference video's expert. We expect that learners will learn the dance movements efficiently by practicing while watching the deepfake video of themselves dancing well. We conducted a user study using 20 participants into two groups: one presented with a deepfake video of the participants dancing, and the other presented with one of the experts dancing. Based on the results of the user study, we provide design implications for a dance practice support system that utilizes deep learning techniques for image generation.

2 RELATED WORKS

2.1 Video self-modeling

The method of improving one's behavior by watching a video of oneself performing the ideal movement is called video self-modeling and has been reported in various settings [6, 9, 10, 26, 29, 34, 37]. Steel et al. [28] applied video self-modeling to control an affected limb in 18 stroke patients. By reversing the video of the patient successfully performing the task on the leg opposite to the affected leg, they created and presented a video in which the patient's affected leg was performing the task normally. Thus, the forward movement of the affected leg was improved, along with the patients' confidence, self-consciousness, and sense of well-being. Diane et al. [27] used video editing to create and present learners with a video of a trampoline performance in which the learner was performing at a higher skill level than their current ability. The result showed improved movements when using the edited video compared to verbal instructions. Although there are problems in video viewing versus other learning approaches, the results indicate that video self-modeling can positively influence teaching movement skills. In our proposed method, we constructed a deep learning model to automatically generate the deepfake videos for video self-modeling, which teaches movement skills effectively.

Several methods have been proposed to apply video self-modeling interactively. Nakanishi et al. [21] proposed a method to support shadowing techniques in a second language by generating and presenting an intermediate video of the mouth movements of the learner and teacher. The evaluation results show that pronunciation is better when the

proposed method is used than when using the video alone. Michael et al. [18] proposed a VR exercise game with an exercise bike that records all past plays and allows the player to race against their future "ghost" self. After a four-week experiment, they found that the method improved physical performance, intrinsic motivation, and flow. Though we have not yet achieved interactive video self-modeling, we believe that in the future, our system will generate and present videos in real-time according to learners' skill levels. As a first step toward realizing such technology, this study examined how the generated deepfake video affects learning dance movements.

2.2 Improving performance by artificial success experiences

Several studies have attempted to improve learners' performance using success experiences. Aymerich-Franch et al. [2] investigated anxiety by showing a doppelganger giving a speech or imagining a successful speech before the actual speech. The results showed that the group of men who saw their doppelganger had reduced anxiety than the group of women who saw their doppelganger. Futami et al. [13] proposed a system to induce a positive mental state and improve performance using a repeated stimulus that is presented when the user succeeds in throwing a dart at a target. The experimental results indicate that associating the success information before the performance improved the user's mental state and performance. Tagami et al. [32] focused on putting golf and developed a virtual golf simulator called *Routine++*, which provides users with comfortable feedback on whether they have put the ball in the hole successfully. They confirmed that *Routine++* improves the novice golfers' performance when playing under pressure. In this research, we aim to improve learning efficiency by showing learners their dancing, which is similar to presenting an artificial success. We believe that our study is novel in its application to dances with complex movements.

2.3 Supporting the learning of dance movements

Many systems have been proposed to support the learning of dance movements, as shown in the survey report by Raheb et al. [23]. Various perceptual methods have been proposed to support learning dance movements, such as tactile presentation using vibration motors [3, 20, 22, 38], auditory presentation using sound [30, 33, 40], and visual presentation using images with various information [14, 15, 24]. In this study, we focus on video viewing using visual perception.

Several learning methods allow learners to look at dance from different angles [41]. In addition, it is effective to practice dance movements facing a mirror [8]. Therefore, many systems that imitate or extend mirrors have been proposed [1, 17, 19]. Choi et al. [5] proposed a system that inverted and displayed video images taken from a camera on the screen as a mirror. A reference dance video was displayed in a separate window, on which the learner's skeleton information was mapped in real-time. In this study, the left half of the screen displays the inverted camera image in real-time, and the right half displays the reference dance video (original and deepfake videos, respectively).

3 DEEPFAKE VIDEO GENERATION

Although we have seen some examples of Deepfake applied to dance [31], no applications have been applied to learning and mastering dance. In this study, we utilized a video generation model based on Chan et al.'s method [4] to generate deepfake videos that serve as references for dance movements (Fig. 1). Our proposed system generates videos by using the skeleton information extracted via OpenPose [42] as an intermediate representation. By normalizing the skeleton information, the system absorbs the differences in body size and position between the learner and dancer in the input and the reference videos, respectively. The system uses a 3-minute video of a learner practicing the dance movements. As an output video, the system generates a video of the learners dancing with the same length and dance movements as

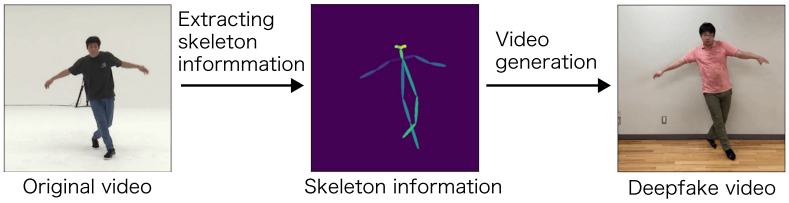

Fig. 1. Flow of deepfake video generation.

the expert dancer in the reference video. Using a cloud service (Amazon EC2) with a Tesla V100 GPU (16 GB), 8 vGPUs, and 61 GB of memory, we trained the model in 40 epochs. Completing the process took about 20 h per input video.

We compared the output videos from the deep learning model between input where learners move freely and where learners imitate the dance movements according to the reference video. Hence, the output video was of higher quality when we input the video in which the learner repeated the same dance movements. Next, we trained the model using several reference videos containing various dance movements and 3-minute input videos of the learner practicing dance movements, from which the deepfake videos are generated. The quality of the deepfake videos of learners dancing was lower when the dance movements included depth, up-and-down, backward-looking, and quick and complicated movements. Specifically, when the reference video included depth movements, the deepfake video could not represent these movements accurately because the skeleton information in the intermediate representation was two-dimensional. When the reference video included up-and-down movements, the learner's position in the video moved slightly up and down. The deepfake video sometimes displayed the learner's face instead of the head's back in the backward-looking movements. In quick and complicated movements, the learner's arm suddenly disappeared or its position seemed to warp. Therefore, we should use the reference video that omits these motions in the user study.

4 USER STUDY

We conducted a user study to verify whether watching videos of themselves performing an expert dancer's movements teaches movement skills effectively. The participants were 20 university students in their 20s (19 males and one female). Their dance experience was limited to junior high school and high school classes and group dance performances at school events such as school festivals. The user study was conducted in a studio with a wooden floor of 43m^2 set up in our university.

4.1 Learning target

Participants learn three dance movements (see figure 2, right) trimmed from videos in the AIST Dance Video Database [35] for use as reference videos. The dance genres are Break, Pop, and Pop having video lengths of 5.9, 9.6, and 8.6 s, respectively; the BPMs of the musical pieces in the videos are 80, 100, and 110; and the names of the videos are gBR_sBM_c01_d04_mBR0_ch01, gPO_sBM_c01_d10_mPO2_ch10, and gPO_sFM_c01_d11_mPO3_ch11, referred to as Dance 1, 2, and 3, respectively, in this paper. Based on the knowledge obtained through the preliminary experiments, to improve the quality of the generated deepfake videos, we selected dance movements on a plane perpendicular to the direction of the camera shooting, with few depths, up-and-down, no looking backward, and no quick and complicated movements. This study was approved by the Ethics Review Board of the University.

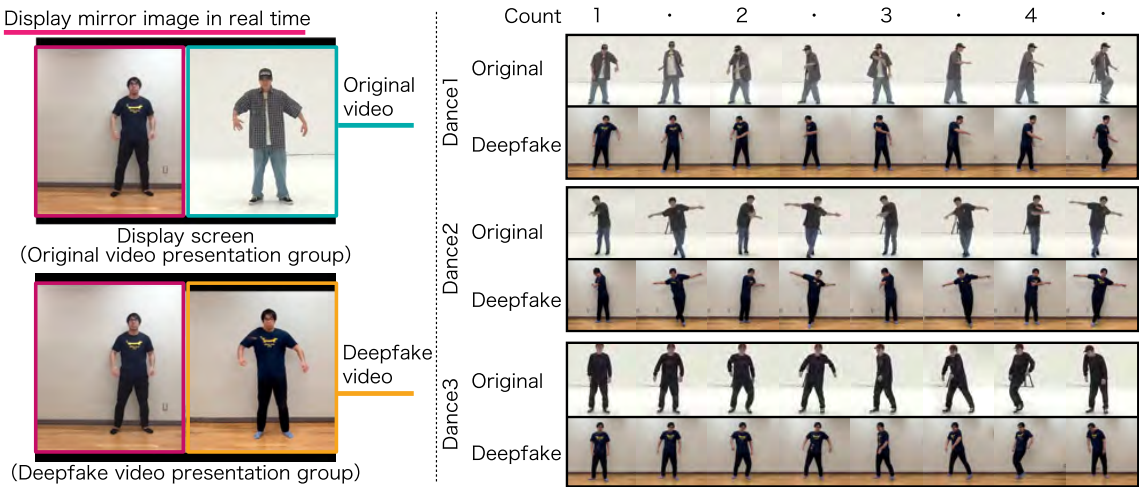

Fig. 2. (Left) Display screen for each group. (Right) Examples of three dance movements.

4.2 Experimental condition

The 20 participants were divided into two equal groups, with one group presented a deepfake video and the other the original video. In the deepfake video presentation group, the participants practice while watching reference videos of themselves performing an expert dancer's movements. Our system generated these videos from the footage taken during practice (PreTraining in the 4.3 Section). In the original video presentation group, the participants practice while watching the reference video of an expert dancer. For aligning the participants' dance proficiency among the groups, the groups were divided based on the dynamic time warping (DTW) cost values calculated in the PreTest described in Section 4.4. The difference in the average DTW cost values in the PreTest was within 10. The average DTW cost of the deepfake presentation group was 814.5 with a standard deviation of 61.5, and that of the original presentation group was 815.2 with a standard deviation of 61.9. As shown in Figure 2, our application's screen displays a mirror image of the participants' image taken through a camera mounted on a PC in the left half, and the right half is a reference video. The mirror image is displayed in real-time. We developed the application with openFrameworks v0.11.2 on macOS Big Sur.

The order in which the participants learned the three dance movements (from Dance 1 to 3) was counterbalanced to avoid overlaps among the participants. The participants practiced while watching the videos on an external display ($90\,cm \times 50\,cm$). The experimental environment is shown in Figure 3. The camera and PC used were an iPhone SE (2nd generation) and a MacBook Pro (13-inch, 2019, Four Thunderbolt 3 ports), respectively. A web browser displayed a timer on the PC screen to indicate the time remaining for learning.

4.3 Experimental procedure

The experimental procedure is shown in Figure 4. The user study was carried out over three days: DAY 1, 2, and 3.

On DAY 1, the participants were briefed on the three dance movements by the experimenter (from Dance 1 to Dance 3). The experimenter explained the application used during the practice and demonstrated the practice to the participants. Next, the participants were informed of a PreTest conducted immediately after the five-minute practice.

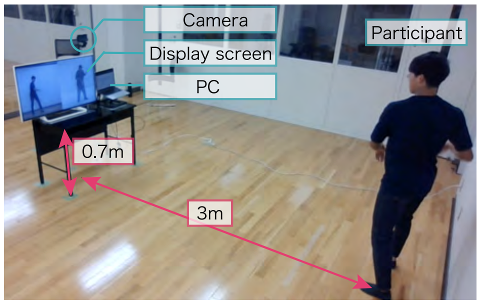

Fig. 3. Experimental environment.

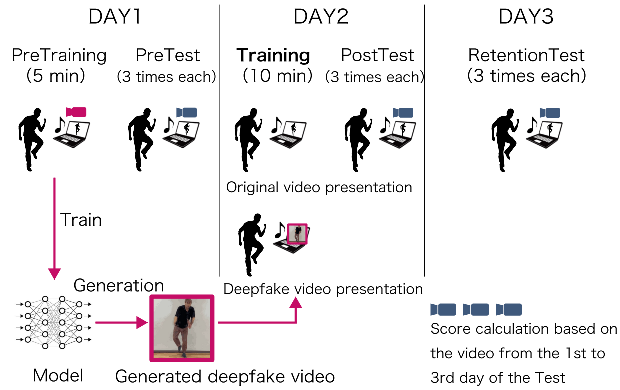

Fig. 4. Experimental process.

The participants were then made to start practicing the dance movements while watching reference videos of the expert dancer through the application. In the PreTest, the participants performed the dance movements three times at intervals of approximately 10 seconds while watching the reference video of the expert dancer while the experimenter filmed them. Day 1 ended when the participant practiced all the three dance movements and completed the PreTest (approximately 20 min). After the end of the experiment, we trimmed 3 min from 1 min after the start of the video and input it into our system's model to start learning. After approximately 20 h, the system generated the deepfake video (used for the deepfake video presentation group) of the participant performing the expert dancer's movements in the reference video.

On DAY 2, the participants practiced the same dance movements and took a PostTest as in DAY 1, but the practice time was set to 10 min. The deepfake video presentation group practiced the dance movements while watching the deepfake video generated from the model. In the PostTest after the practice, both groups performed the dance movements while watching the reference video of the expert dancer three times at intervals, with the experimenter again filming them. After the PostTest, the participants were made to answer a questionnaire rating the difficulty of learning each dance movement on a 7-point scale (1: very easy to 7: very difficult. To determine the participants' self-efficacy, they were asked the statement *"I think I can learn to dance if I keep practicing."* for each dance movement on a 5-point Likert scale (1: strongly disagree to 5: strongly agree). To know the participants' sense of accomplishment, they were asked the statement *"I could master the dance."* for each dance movement on a 5-point Likert scale (1: strongly disagree to 5: strongly agree). The experimenter provided a free writing column for other comments on the user study, and the participants were asked to answer there. Moreover, only the deepfake video presentation group was asked to answer the statement *"I felt as if I were dancing."* for each dance movement on a 5-point Likert scale (1: strongly disagree to 5: strongly agree). In addition, the experimenter asked the participants to respond to a free-description questionnaire for comments on their sensations when they viewed the deepfake video. The responses, gathered via Google forms, were then translated by the first author. Day 2 ended after answering the questionnaire (approximately 45 min).

On DAY 3, the experimenter conducted a RetentionTest more than 20 h after DAY 2 ended to confirm whether the effect of the practice was retained. All participants performed the dance movements three times at intervals while watching the reference video of the expert dancer, and the experimenter filmed them. Day 3 ended when the RetentionTest of all dance movements was completed (about 5 min).

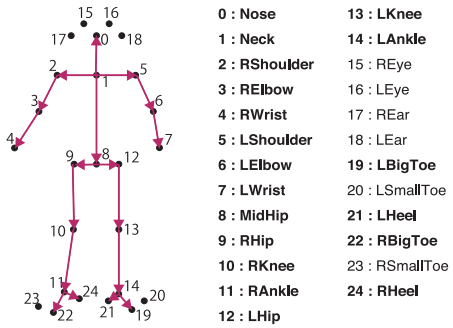

Fig. 5. Vectors used for calculation from the skeleton information extracted by OpenPose.

4.4 Evaluation method

We calculated the cost of the difference in movements between the practice and reference videos during the three tests. First, we calculated the cross-correlation function between the audio in the reference video and used musical pieces. We trimmed the videos to the point where the cross-correlation function value is the highest. Using this as the start time, we aligned the start time of all the dance movements. Next, we applied OpenPose [42] to the reference and trimmed videos to calculate 25 feature points (2-dimensional x, y coordinates) as shown in Figure 5.

Considering the differences in body size of the participants, we set the 18 skeletons represented by the red arrows in Figure 5 as vectors. We then converted each vector into a unit vector. A confidence level (0 to 1) was calculated for each feature point. The vectors with confidence level of at least 0.1 for each connected feature point were extracted, while 0 was assigned to the remaining vectors 36-dimensional vectors per frame were obtained. These 36-dimensional vectors are the feature vectors per video, and DTW is calculated for the feature vectors between participants and reference videos using the FastDTW library [25] to calculate the distance. This distance was used as the DTW cost to evaluate how well the dancers danced to the reference video. The lower the DTW cost, the higher the similarity between the reference and participant's dance movements. In addition, because OpenPose estimates skeletons on two-dimensional coordinates, the movements in the depth direction are ignored. However, we assume that this is not a problem for use as an evaluation index. In addition, because each test used a reference video that the system played repeatedly, many participants started moving later than the exact start timing of the dance movements. Therefore, we skipped the first 60 frames, which correspond to about 1 s, and used the feature vectors from the 61st frame when calculating the DTW cost.

5 RESULT

The average DTW costs for Dance 1 to Dance 3 are shown in Figure 6. The error bars indicate the standard error. We conducted a two-factor between-subjects ANOVA for DTW cost, with the schedule (DAY 1 to DAY 3) and presentation group as factors. The result showed a significant difference in the schedule ($F(2, 174) = 5.75, p < 0.01$), while no significant differences were observed in the presentation group. We also conducted multiple comparisons using the Holm method, which revealed significant differences between the scores of DAY 1 and DAY 2, and between DAY 1 and DAY 3 ($DAY\ 1 > DAY\ 2$, $DAY\ 1 > DAY\ 3$, $p < 0.05$). The average DTW cost for each type of dance is shown in Figure 7. The error bars indicate the standard error. We conducted a two-factor between-subjects ANOVA on the DTW cost

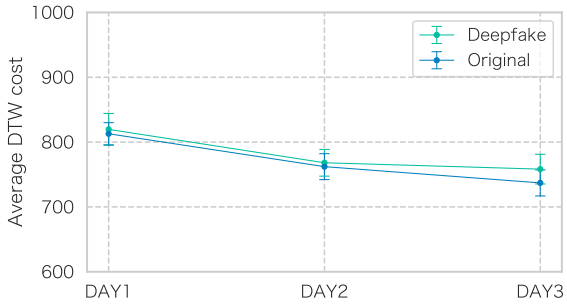

Fig. 6. Average DTW costs for Dance 1 to Dance 3

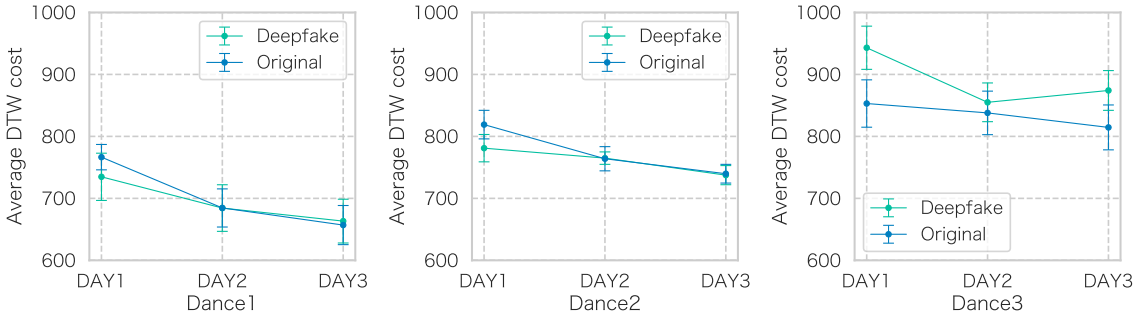

Fig. 7. Average DTW cost per dance

of Dance 1, with the schedule and presentation group as factors. The results showed a significant difference for the schedule ($F(2, 54) = 4.07, p < 0.05$) but no significant differences for the presentation group. Multiple comparisons using the Holm method showed a significant difference between DAY 1 and DAY 3 ($DAY\ 1 > DAY\ 3$, $p < 0.05$). of the above analysis was repeated for Dance 2 and 3; multiple comparisons revealed a significant difference between DAY 1 and DAY 3 ($DAY\ 1 > DAY\ 3$, $p < 0.05$) in Dance 2, while no significant differences were observed between the two factors in Dance 3.

Next, we reviewed the questionnaire results. Figure 8 shows the results of the questionnaire on the difficulty in learning the dance movements; the error bars indicate the standard errors. A two-factor between-subjects ANOVA with the presentation group and type of dance as factors showed a significant difference between the types of dance ($F(2, 54) = 73.50, p < .01$). Multiple comparisons using the Holm method showed significant differences in all combinations of dances ($p < .05$). Therefore, we conclude that Dance 1, 3, and 2 are progressively more difficult dance movements. In contrast, there was no significant difference among the presentation groups.

Figure 9 shows the results of the responses on *"I think I can learn to dance if I keep practicing."* The results indicate the self-efficacy of the participants. The two-factor between-subjects ANOVA showed a significant differences in the presentation group ($F(1, 54) = 3.904, p < 0.1$) and type of dance ($F(2, 54) = 61.51, p < 0.01$). As for the presentation groups, the mean 2.8 of the deepfake video presentation group is significantly smaller than the mean 3.2 of the original video presentation group. Multiple comparisons using the Holm method showed significant differences in all combinations of dances ($p < 0.05$). Thus, the participants ranked the dances in the order Dance 1, 3, and 2 in terms of

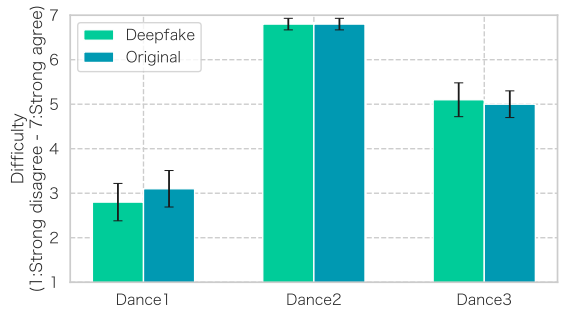

Fig. 8. Participants' answers to *"I think I can learn to dance if I keep practicing."*

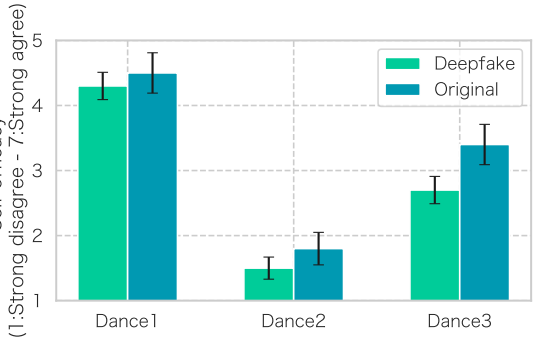

Fig. 9. Participants' answers to "I think I can learn to dance if I continue to practice."

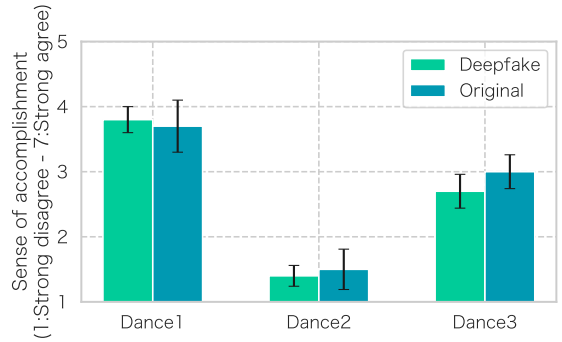

Fig. 10. Participants' answers to *"I could master the dance."*

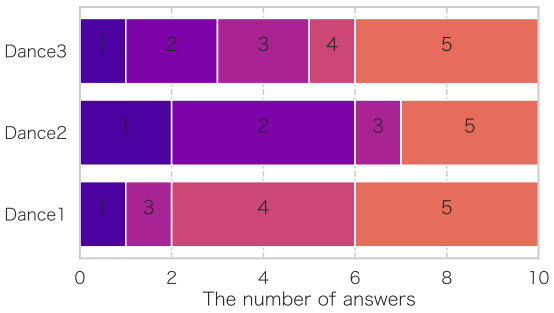

Fig. 11. Participants' answers to *"I felt as if I were dancing."*

ease of learning. In contrast, there is no significant difference among the presentation groups for Dance 3, although the mean difference is above 0.5.

Figure 10 shows the results of the responses on *"I could master the dance,"* indicating the participants' sense of accomplishment. A two-factor between-subjects ANOVA with the presentation group and type of dance as factors showed significant differences between the types of dance ($F(2, 54) = 35.65, p < .01$). Multiple comparisons using the Holm method showed significant differences for all combinations of dances ($p < .05$), and that Dance 1 had a higher sense of accomplishment while Dance 2 had the lowest. In contrast, there are no significant differences between the scores of the presentation groups. Figure 11 shows the results of the questionnaire *"I felt as if I were dancing"* rated on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree) by the deepfake video presentation group. We conclude that the easier the dance movements, the more it made the participants feel as if they were dancing.

The participants' comments in the comment field of the questionnaire had some positive comments on understanding movements and how to move. The comments are as follows: *"I was able to see myself dancing, so it was easy to know how to move," "In the reference video, I cannot understand what kind of movement the dancer was doing. In the deepfake video, the complex movements seemed to be a little easier," "I thought I was able to notice more differences between my movements and the dancer in the PreTraining," "I thought it was easy to understand the timing of the movements," "Since the deepfake*

and the current image of myself were shown side by side, it was easier to compare and correct my movements than with the dancer image," "I felt that it was easier to compare the movements of each part of the body because the deepfake had the same body shape." These comments indicate the ability of the deepfake video in improving the understanding of movements.

Additional comments such as *"It was a strange feeling because it was a video of myself doing a movement that I should not have been able to do, but it was easy to visualize the movement in my brain," "I was moving my body thinking that I was dancing like in the deepfake video. Sometimes I looked at the mirror image of myself and compared it with the deepfake, and I noticed the points where I was not moving well,"* indicating the possibility of supporting the movement.

The ability of deepsake to increase learner motivation is found in comments such as *"My motivation went up because I could see myself getting better."* and *"I could see myself dancing well in the video, so I can enjoy practicing with the illusion that I am dancing well."*

However, there were also some negative comments. *"There were some noises in the deepfake video compared with the reference video, so that I could not understand some details of the movements," "In the video of Dance 1, it was difficult to figure out which foot was in front of another one. If the dance movements contain difficult parts, even if using a deepfake video, it was difficult to imagine it," "The quality of the image was not very good, so it was difficult to see the detailed movements of the fingers."* Most of these comments were related to the low image quality generation, and will be improved in the future.

A considerable variation in the answers is observed to the question, *"I felt as if I were dancing."* Because video self-modeling is more effective in making the participants feel like they were dancing, the results may change based on the answers to the questionnaire. Therefore, we check the average DTW cost trend based on the questionnaire responses. Specifically, we extracted the participants who answered 1 or 5 for all dances and checked their learning results. The results showed that participant (A), who did not feel as if dancing, did not learn well, while participants (B, C, and D), who gave high ratings, tended to learn relatively well (Fig. 12). Though the number of participants who showed this tendency was small, further investigation is required.

6 DISCUSSION

The results of the average DTW costs for Dance 1 to 3 in Figure 6 show no significant differences between the groups, although we could see it contributed to the participants' learning. One possible reason is that the low quality of the generated video, as seen in the participants' comments, may have affected the participants' learning. Hence, generating a system with higher quality images that remove the unnaturalness of the images is required. In addition, there was no significant difference between the groups for each type of dance. However, the role of dancing difficulty levels in these videos needs further investigation. In addition, the small screen size ($90\,cm \times 50\,cm$) may have diminished its role as a mirror and prevented the effect of self-modeling. Using a relatively large display has been shown to reduce the error rate of posture guidance by Elsayed et al. [11],. In addition, the benefits included in the deepfake video obtained in the participants' comments on the questionnaire may appear as quantitatively measurable effects when using a larger display.

The results of Figure 9 show that self-efficacy tended to be significantly lower in the deepfake video presentation group than in the original group, indicating that the former may reduce self-efficacy. The following comments support the above statement. *"I got the illusion that I can dance as well as in the reference video. Therefore, I felt that I frequently had to compare my movements with the mirror images,"* and *"I felt strange because I could dance even though I could not actually dance."* We suspect that the deepfake video presentation of actions that are not immediately attainable at the

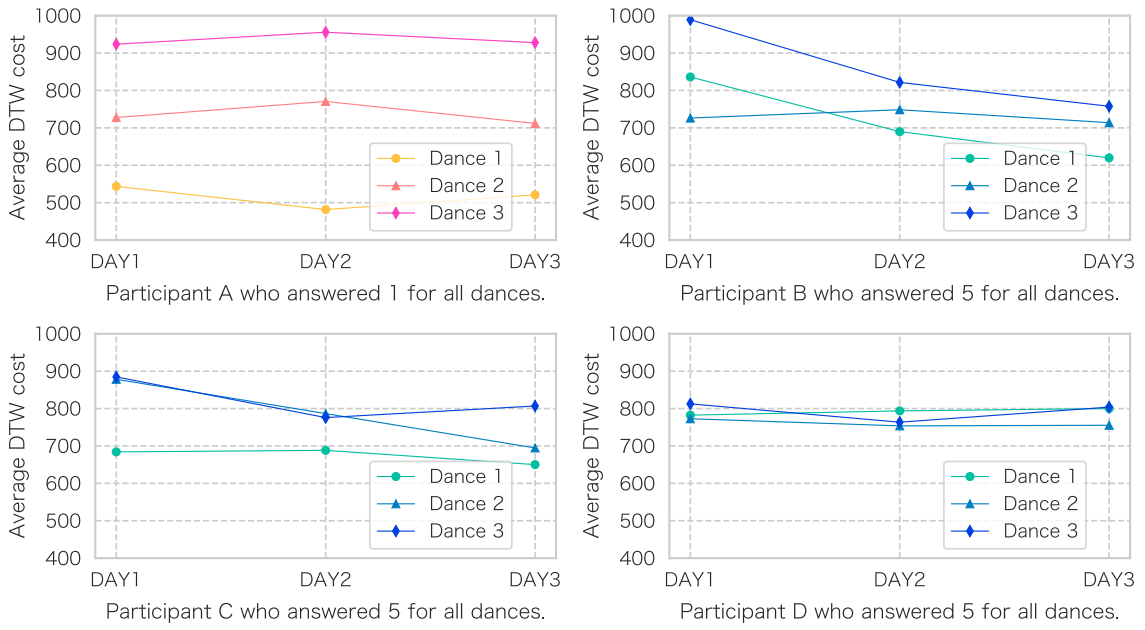

Fig. 12. Average DTW costs per participant based on *"I felt as if I were dancing."* responses.

current skill level may have reduced the score of self-efficacy. On the other hand, there were some comments such as *"I felt a little uncomfortable, but I felt as if I was doing a good dance with my own face, and I felt that I might be able to dance this movement well"* and *"It seemed impossible to imitate the original version video that dancers were doing, but when I saw myself dancing in the deepfake video, I felt that I might be able to do so."* Some individual characteristics may enhance self-efficacy, and we believe that further research is needed.

Our proposed system is mainly intended for learning street dance and would be particularly effective in copying popular dance styles. However, our system does not consider the large differences in skeletal structure. The system may fail for children who want to learn dance due to the differences in skeletal structure between adults and children and needs continued investigation.

7 DESIGN IMPLICATION

In the results shown in Fig. 11, the responses varied, with some participants saying that they did not feel as if they were dancing upon watching the generated images. Such comments include *"The body of the deepfake is small," "Sometimes I could not see the toes," "I felt that the direction of the face was not natural,"* and *"I feel that the deepfake body size is not correct."* Video self-modeling is based on the premise of creating the illusion that oneself are performing the movements. To obtain the effects of video self-modeling, enhancing the sense that one is performing the movements is desirable. Therefore, it is necessary to eliminate the unnaturalness described above by choosing video generation methods [16] with more accurate image quality. The misalignment can be eliminated by applying a process that aligns the person's position in the generated deepfake video with that in the reference video. In addition, by focusing on each body part, such as the hand, we can create and apply a model that allows checking the quality of each part to regenerate the faulty parts.

The results in Figures 8 and 11 show that the more difficult the learner perceives the dance movements to be, the weaker the feeling that they are dancing. In addition to methods such as adjusting the opponents [18] and tools [36] according to the learners' skill level, it is advisable to present deepfake videos according to the learners' level of dance movements. For example, instead of creating a deepfake video applied to an expert's dance movements, the system generates a deepfake video that is one step closer to an expert dancer. In other words, by generating and presenting intermediate-level dance movements between novices and experts, the learners can practice while referring to dance movements that are much ahead of their dance level. We believe that we need to develop technologies that can morph or blend the skills of expert dancers and those of beginners.

8 SUMMARY

We propose a learning support system that allows learners to learn dance movements by watching a deepfake video of themselves mastering an expert dancer's movements. A user study with 20 participants verified the effectiveness of the generated deepfake video for dance learning. The results showed no significant differences compared to watching the reference video. The questionnaire results showed that the self-efficacy of the group presented with the deepfake video was significantly lower. Based on the results, we provide design implications for a dance practice support system using deep learning techniques for image generation.

ACKNOWLEDGMENTS

This work was supported by Softbank Corp. and JST, CREST Grant Number JPMJCR18A3, Japan.

REFERENCES

[1] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: Enhancing Movement Training with an Augmented Reality Mirror. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, UIST '13*. 311–320.

[2] Laura Aymerich-Franch and J Bailenson. 2014. The Use of Doppelgangers in Virtual Reality to Treat Public Speaking Anxiety: a Gender Comparison. In *Proceedings of the International Society for Presence Research Annual Conference*. 173–186.

[3] Hector Camarillo-Abad, Alfredo Sánchez, Oleg Starostenko, and Maria Sandoval Esquivel. 2019. A Basic Tactile Language to Support Leader-Follower Dancing. *Journal of Intelligent & Fuzzy Systems* 36 (2019), 5011–5022.

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. 2019. Everybody Dance Now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV '19*.

[5] Jong-Hyeok Choi, Jae-Jun Lee, and Aziz Nasridinov. 2021. Dance Self-Learning Application and Its Dance Pose Evaluations. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing, SAC '21*. 1037–1045.

[6] Shannon E. Clark and Diane M. Ste-Marie. 2007. The Impact of Self-as-a-model Interventions on Children's Self-regulation of Learning and Swimming Performance. *Journal of Sports Sciences* 25, 5 (2007), 577–586.

[7] Thomas L Creer and Donald R Miklich. 1970. The Application of a Self-modeling procedure to Modify Inappropriate Behavior: A Preliminary Report. *Behaviour Research and Therapy* 8, 1 (1970), 91–92.

[8] Karen Dearborn and Rachael Ross. 2006. Dance Learning and the Mirror: Comparison Study of Dance Phrase Learning with and without Mirrors. *Journal of Dance Education* 6 (2006), 109–115.

[9] Peter Dowrick. 2012. Self Model Theory: Learning From the Future. *Wiley Interdisciplinary Reviews: Cognitive Science* 3 (2012), 215–230.

[10] Peter Dowrick. 2012. Self Modeling: Expanding the Theories of Learning. *Journal of Psychology in the Schools* 49 (2012), 30–41.

[11] Hesham Elsayed, Philipp Hoffmann, Sebastian Günther, Martin Schmitz, Martin Weigel, Max Mühlhäuser, and Florian Müller. 2021. CameraReady: Assessing the Influence of Display Types and Visualizations on Posture Guidance. In *Designing Interactive Systems Conference 2021, DIS '21*. 1046–1055.

[12] Minoru Fujimoto, Tsutomu Terada, and Masahiko Tsukamoto. 2012. A Dance Training System that Maps Self-Images onto an Instruction Video. In *Proceedings of the 5th International Conference on Advances in Computer-Human Interactions, ACHI '12*. 309–314.

[13] Kyosuke Futami, Tsutomu Terada, and Masahiko Tsukamoto. 2016. Success Imprinter: A Method for Controlling Mental Preparedness Using Psychological Conditioned Information. In *Proceedings of the 7th Augmented Human International Conference 2016, AH '16*. Article 11, 8 pages.

[14] Chan Jacky C.P., Leung Howard, Tang Jeff K.T., and Komura Taku. 2011. A Virtual Reality Dance Training System Using Motion Capture Technology. *Journal of IEEE Transactions on Learning Technologies* 4, 2 (2011), 187–195.

[15] Matthew Kyan, Guoyu Sun, Haiyan Li, Ling Zhong, Paisarn Muneesawang, Nan Dong, Bruce Elder, and Ling Guan. 2015. An Approach to Ballet Dance Training through MS Kinect and Visualization in a CAVE Virtual Reality Environment. *Journal of ACM Transactions on Intelligent Systems and Technology* 6, 2, Article 23 (2015), 37 pages.

[16] Wen Liu, Zhixin Piao, Min Jie, Wenhan Luo, Lin Ma, and Shenghua Gao. 2019. Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV '19.*

[17] Zoe Marquardt, João Beira, Natalia Em, Isabel Paiva, and Sebastian Kox. 2012. Super Mirror: A Kinect Interface for Ballet Dancers. In *Extended Abstracts Proceedings of the 2012 Conference on Human Factors in Computing Systems, CHI EA '12.* 1619–1624.

[18] Alexander Michael and Christof Lutteroth. 2020. Race Yourselves: A Longitudinal Exploration of Self-Competition Between Past, Present, and Future Performances in a VR Exergame. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20.* 1—-17.

[19] Luis Molina-Tanco, Carmen García-Berdonés, and Arcadio Reyes-Lecuona. 2017. The Delay Mirror: A Technological Innovation Specific to the Dance Studio. In *Proceedings of the 4th International Conference on Movement Computing, MOCO '17.* Article 9, 6 pages.

[20] A. Nakamura, S. Tabata, T. Ueda, S. Kiyofuji, and Y. Kuno. 2005. Dance Training System with Active Vibro-Devices and a Mobile Image Display. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS '05.* 3075–3080.

[21] Yoko Nakanishi and Yasuto Nakanishi. 2015. Use of an Intermediate Face between a Learner and a Teacher in Second Language Learning with Shadowing. In *Proceedings of the 6th Augmented Human International Conference, AH '15.* 113—-116.

[22] Thomas Peeters, Eric van breda, Wim Saeys, Evi Schaerlaken, Jochen Vleugels, Steven Truijen, and Stijn Verwulgen. 2019. Vibrotactile Feedback During Physical Exercise: Perception of Vibrotactile Cues in Cycling. *International Journal of Sports Medicine* 40 (04 2019).

[23] Katerina El Raheb, Marina Stergiou, Akrivi Katifori, and Yannis Ioannidis. 2019. Dance Interactive Learning Systems: A Study on Interaction Workflow and Teaching Approaches. *Journal of ACM Computing Surveys* 52, 3, Article 50 (2019), 37 pages.

[24] Jean-Philippe Rivière, Sarah Fdili Alaoui, Baptiste Caramiaux, and Wendy E. Mackay. 2019. Capturing Movement Decomposition to Support Learning and Teaching in Contemporary Dance. *Proceedings of the ACM on Human-Computer Interaction, PACM HCI '19* 3, CSCW, Article 86, 22 pages.

[25] Stan Salvador and Philip Chan. 2007. Toward Accurate Dynamic Time Warping in Linear Time and Space. *Jornal of Intelligent Data Analysis* 11, 5 (2007), 561–580.

[26] Joanna Starek and Penny Dorrance McCullagh. 1999. The Effect of Self-modeling on the Performance of Beginning Swimmers. *Journal of Sport Psychologist* 13 (1999), 269–287.

[27] Diane Ste-Marie, Kelly Vertes, Amanda Rymal, and Rose Martini. 2011. Feedforward Self-Modeling Enhances Skill Acquisition in Children Learning Trampoline Skills. *Journal of Frontiers in Psychology* 2 (2011), 155.

[28] Kylie Steel, Kurt Mudie, Remi Sandoval, David Anderson, Sera Dogramaci, Mohammad Rehmanjan, and Ingvars Birznieks. 2017. Can Video Self-Modeling Improve Affected Limb Reach and Grasp Ability in Stroke Patients? *Journal of motor behavior* 50 (2017), 1–10.

[29] Kylie A Steel, Roger Adams, Susan Coulson, Colleen G. Canning, and Holly J Hawtin. 2013. Video Self-model Training of Punt Kicking. *International Journal of Sport and Health Science* 11 (2013), 49–53.

[30] Landry Steven and Jeon Myounghoon. 2020. Interactive Sonification Strategies for the Motion and Emotion of Dance Performances. *Journal of Multimodal User Interfaces* 14 (2020), 167–186 pages.

[31] sway. 2019. *Sway.* https://getsway.app/

[32] Shoichi Tagami, Shigeo Yoshida, Nami Ogawa, Takuji Narumi, Tomohiro Tanikawa, and Michitaka Hirose. 2017. Routine++: Implementing Pre-Performance Routine in a Short Time with an Artificial Success Simulator. In *Proceedings of the 8th Augmented Human International Conference, AH '17.* Article 18, 9 pages.

[33] Grosshauser Tobias, Bläsing Bettina, Spieth Carinna, and Hermann Thomas. 2012. Wearable Sensor-Based Real-Time Sonification of Motion and Foot Pressure in Dance Teaching and Training. *Journal of the Audio Engineering Society* 60, 7/8 (2012), 580–589.

[34] Jill Tracey. 2011. Benefits and Usefulness of a Personal Motivation Video: A Case Study of a Professional Mountain Bike Racer. *Journal of Applied Sport Psychology* 23, 3 (2011), 308–325.

[35] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. 2019. AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR '19.* 501–510.

[36] Dishita G Turakhia, Andrew Wong, Yini Qi, Lotta-Gili Blumberg, Yoonji Kim, and Stefanie Mueller. 2021. Adapt2Learn: A Toolkit for Configuring the Learning Algorithm for Adaptive Physical Tools for Motor-Skill Learning. In *Designing Interactive Systems Conference 2021, DIS '21.* 1301–1312.

[37] Kelly A. Vertes and Diane M. Ste-Marie. 2013. Trampolinists' Self-controlled Use of a Feedforward Self-modeling Video in Competition. *Journal of Applied Sport Psychology* 25, 4 (2013), 463–477.

[38] Steeven Villa, Jasmin Niess, Bettina Eska, Albrecht Schmidt, and Tonja-Katrin Machulla. 2021. Assisting Motor Skill Transfer for Dance StudentsUsing Wearable Feedback. In *2021 International Symposium on Wearable Computers, ISWC '21.* 38–42.

[39] Paul Ward, Nicola J Hodges, and A Mark Williams. 2004. Deliberate Practice and Expert Performance: Defining The Path to Excellence. In *Journal of Skill acquisition in sport.* 255–282.

[40] Tomoyuki Yamaguchi and Hideki Kadone. 2014. Supporting creative dance performance by grasping-type musical interface. In *Proceedings of the 2014 IEEE International Conference on Robotics and Biomimetics, ROBIO '14.* 919–924.

[41] Shuo Yan, Gangyi Ding, Zheng Guan, Ningxiao Sun, Hongsong Li, and Longfei Zhang. 2015. OutsideMe: Augmenting Dancer's External Self-Image by Using A Mixed Reality System. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems,*

CHI EA '15. 965–970.

[42] Cao Zhe, Hidalgo Gines, Simon Tomas, Wei Shih-En, and Sheikh Yaser. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).