VoLearn: A Cross-Modal Operable Motion-Learning System Combined with Virtual Avatar and Auditory Feedback

CHENGSHUO XIA, Keio University, Japan
XINRUI FANG, Keio University, Japan
RIKU ARAKAWA, Carnegie Mellon University, United States
YUTA SUGIURA, Keio University, Japan

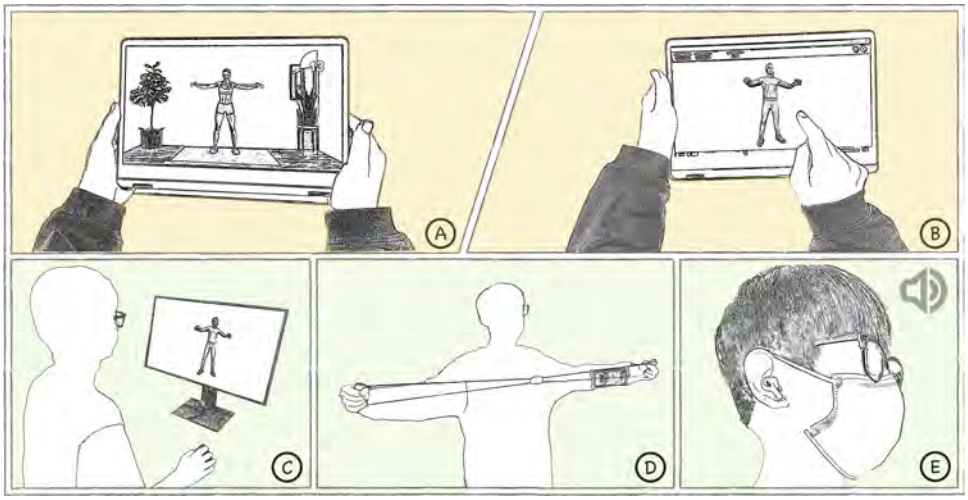

Fig. 1. *VoLearn* helps the user in motion learning by operating the motion animation converted from a video source and receiving auditory feedback. It can be used in different contexts. One can design the motion from the video as a trainer. The other trainee studies the motion with the virtual avatar and auditory feedback. (a) The desired motion video. (b) Motion editing interface to design the motion. (c) User-end motion. (d) The user wears the smartphone. (e) The auditory feedback helps the user reduce the motion amplitude and time errors.

Conventional motion tutorials rely mainly on a predefined motion and vision-based feedback that normally limits the application scenario and requires professional devices. In this paper, we propose *VoLearn*, a cross-modal system that provides operability for user-defined motion learning. The system supports the ability to import a desired motion from RGB video and animates the motion in a 3D virtual environment. We built an interface to operate on the input motion, such as controlling the speed, and the amplitude of limbs for the respective directions. With exporting of virtual rotation data, a user can employ

Authors' addresses: Chengshuo Xia, csxia@keio.jp, Keio University, Yokohama, Japan; Xinrui Fang, xinrui.fang@keio.jp, Keio University, Yokohama, Japan; Riku Arakawa, Carnegie Mellon University, Pittsburgh, United States; Yuta Sugiura, Keio University, Yokohama, Japan.

2474-9567/2022/6-ART81 $15.00
https://doi.org/10.1145/3534576

a daily device (i.e., smartphone) as a wearable device to train and practice the desired motion according to comprehensive auditory feedback, which is able to provide both temporal and amplitude assessment. The user study demonstrated that the system helps reduce the amplitude and time errors of motion learning. The developed motion-learning system maintains the characteristics of high user accessibility, flexibility, and ubiquity in its application.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Cross-modality, feedback, motion learning, virtual avatar

ACM Reference Format:
Chengshuo Xia, Xinrui Fang, Riku Arakawa, and Yuta Sugiura. 2022. VoLearn: A Cross-Modal Operable Motion-Learning System Combined with Virtual Avatar and Auditory Feedback. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 81 (June 2022), 26 pages. https://doi.org/10.1145/3534576

1 INTRODUCTION

The need to learn motion has permeated people's daily lives in multiple ways. Motion learning helps individuals in physical rehabilitation [3], sports training [14], dance [60], exercise [17, 30], and other types of activities [18, 23, 48]. Conventionally, users can realize tele-education to study a new motion through instruction from professional video tutorials. However, reproducing the motion presented in a video is insufficient for effective learning, and individuals vary in their success with it. Thus, considerable effort goes into promoting and developing motion-learning systems with better learning experience [40, 42, 64].

Based on the motion capture (MoCap) technique, a motion learning system for a novice can be achieved in various ways. In terms of commercial tools, specialized equipment, such as Microsoft Kinect [10], can support several interfaces that realize motion guidance according to capture the user's body. In addition, some smartphone-based applications, such as Kaia Trainer [21], Seven [49], and Fitness Builder [46], utilize a built-in camera to capture a user's motion and present vision guidance to help users perform motions.

To better aid the motion-learning process, research in the field of human–computer interaction has focused on various ways to help users understand and perform correct motions. Generally, motion-learning systems are developed in combination with dedicated devices to obtain user movement (e.g., skeleton information obtained from a vision sensor [8] and kinematic data obtained from an inertial measurement unit (IMU) [18]). Although these systems facilitate accurate and professional detection of motion, the required devices and the employed lab-based learning spaces induce considerable burden on the users and limit pervasive applications. Besides, popular solutions that rely on a vision sensor request users to be active within a specific area (e.g., in front of a camera) during usage. In addition to the hardware, most systems are utilized based on a predefined motion, because of which the system exhibits low flexibility and difficulty in adjusting to different usage conditions. Thus, finding a new solution based on an efficient and convenient system, and improving the flexibility of a learning system is able to benefit more motion learning applications and engage more users into it.

Therefore, in this paper, we propose *VoLearn*, which concentrates on two main issues faced by current systems: customized motion editing and an efficient guiding system based on a daily device. *VoLearn* supports motion video input from multiple sources and converts the video to an editable virtual 3D avatar. In a virtual 3D environment, we provide an interface to enable the user to establish a customized motion. A user is able to modify the motion to control the speed, the body movement amplitude, to design a motion. Additionally, a virtual sensor is designed to obtain the rotation data of the virtual avatar's body segment. During the motion learning period, a smartphone worn on the human body detects the real body segment rotation information. The exported virtual sensor data is then compared with the real sensor's rotation data attached to the user's body. The system will use audio to provide instructions and feedback to the user. The main contributions of this paper can be summarized as follows:

(1) A 3D virtual motion system is established based on an RGB video input that supports motion speed adjustment, amplitude modification, and body motion design. In addition, a virtual rotation sensor is designed to

export the virtual avatar's body segment rotation data. The exported rotation information is compared with real sensor data to provide motion instruction as feedback.

(2) An error-based comprehensive feedback system is developed on a smartphone to supply motion instruction based on auditory feedback, which can guide the user in motion learning.

(3) The effect of *VoLearn* on motion learning is validated. The results show that the auditory feedback of *VoLearn* helps a user reduce the amplitude and time errors in motion learning and indicates the system's usability from users' comments.

(4) Informal interviews are conducted with various professional users, such as fitness trainers, clinical nurses, and rehabilitation therapists, under the demonstration of the system. We report the obtained implications about the system's elements and application scenes.

2 RELATED WORK

2.1 Motion Tutorials with Different Modalities

So far, many motion-learning systems have been developed to help novice learners learn motion and have been applied in various fields (e.g., ballet [20, 23, 61], factory operations [7], sports training assistance [15], Tai-Chi Chuan exercises [12, 23, 36] and physical rehabilitation [3, 10, 56, 57]). Considerable efforts have been made to provide lightweight, intuitive, and effective systems for users. Virtual reality (VR) devices enable teaching in a more realistic environment. For example, the "Just Follow Me" system [69] proposed utilizing a ghost metaphor from a first-person perspective in presenting a standard motion trajectory. A similar metaphor and overlaying process were employed in subsequent VR-based systems for the trajectory-based following [28, 29] and remote posture guidance [25].

As a video contains informative resources, intensive efforts have been made to augment traditional video-based tutorials, such as a 3D assembly tutorial [68], perspective translation [44], and machine task completion [7]. Body language is also a good instructor for enhancing the interactivity of watching a video, such as determining the frame for beginning playback in PoseAsQuery [22] and fitting the learning process in ReactiveVideo [13]. The addition of depth information to a user's 3D motion data can enable applications such as an interactive authoring process in YouMove [1], multi-view corrective guidance in Physio@Home [57], golf training [27], and physical rehabilitation motion [10, 59]. Additionally, converting skeleton trajectory information to other representations, such as in Supper Mirror [41], ballet tutorials with a mirror [20, 23], hints from the projection on a user's skin in LightGuide [53], step projection in gait training [48], and upper-limb exercise in SleeveAR [54], is an intuitive solution as well.

In addition to vision-based systems, an IMU is the most frequently adopted device in applications such as knee rehabilitation [3], ergonomic in NurseCare [18], joint angle evaluation for a frozen shoulder exercise [56], weight training anomaly detection [34], and assessing the movement [52].

2.2 Feedback Based on Error

To engage a user and effectively meet the learning goal, several supporting forms for communication between the motion-learning system and users have been investigated (i.e., feedforward and feedback processes) [20, 51]. The feedforward approach provides directional instructions, such as an arrow indication [13], and step guidance in augmented reality [7]. Nevertheless, for mastery, an error-based feedback mechanism can assist in improving motion learning [31, 51].

For whole-body motion, holistic feedback is created generally, as in the collection of whole skeleton information by using a MoCap system [58] and joint vectors through principal component analysis [22]. Specific body segment comparisons are also frequently used, such as in Lightguide [53]. The tactile interaction for kinesthetic learning (TIKL) tracks the joint angle of the upper limb and calculates the error to provide feedback [37]. Gibbons [20]

defined four types of feedback modes (value, corrective, neutral, and ambiguous) in visual, verbal, and kinesthetic channels. A scoring mechanism, such as SleeveAR [54] and an exercise training game [59] can be designed to encourage users to improve their behavior. In addition, Trajkova wt al. [60] combined emoji expressions with a verbal presentation to define a feedback-adapted ballet framework.

Relying on the actuators attached to the user can also be an effective way to improve learning efficiency [40]. Schonauer et al. [47] aimed to correct a user's motion based on visual, pneumatic, and tactile feedback. Moreover, vibrotactile feedback can also be used to remind a user to alter their current position [42]. TIKL [37] employs eight actuators mounted on the skin, and the user receives different vibration intensities according to the error value.

2.3 Designing Auditory Stimulation

To help with the guiding process, audio can be employed to aid users in basic stimulation. The investigation related to audio utilization has been deeply realized in assisting visually impaired people in wayfinding [73], target searching [43], attracting users' attention [2], and reaching tasks [66]. In addition, research on audio design space has focused on the augmentation of other applications, such as those in collaboration with computer vision systems [45], text entry on touchscreens [4], games [39, 67], VR experiences [6], mobile devices [26], gait analysis [9], and visual searching [32]. Audio can be designed in one of the following two ways: through vocal characteristics or through the characteristics of the audio itself (such as audio tone, pitch, rhythm, and musical phrases [45]). These elements can be used metaphorically or as natural expressions as part of a reminding task.

In contrast, audio is less frequently used in the motion-learning process. For example, a chronic obstructive pulmonary disease (COPD) trainer was designed to help COPD patients conduct training exercises [55], and an audio element was used to convey the assessment of the patient's motion. Therefore, the design space and effect of audio elements added to motion-learning systems need to be studied further.

2.4 Cross-modal Approach in Human Activity Systems

Systems that deal with human activity based on machine learning techniques inevitably require considerable data to train the model. To efficiently obtain diverse modalities of human activity data, previous studies have applied cross-modal approaches to generate simulated data [33, 65, 72, 74]. IMUSim [70] for the first time, generated a simulation of an IMU device at a signal level according to a motion trajectory. In addition, IMUTube [35] extracted virtual acceleration data from an RGB video to train a human action recognition (HAR) classifier. Similarly, Liu et al. [38] and Zhang et al. [71] used videos to generate on-body accelerometers and finger inertial sensors, respectively. Previous studies have concentrated on dataset generation to alleviate the task of collecting training samples for classical machine learning in HAR systems. In contrast, our system generates virtual rotation data of the human body from videos in a cross-modal manner. This approach enables a novel function, namely operability, which allows users to control the speed and amplitude of motions. Such variations are important in motion learning and in supporting wider scenarios.

3 DESIGN SPACE FINDINGS

To better position our proposed system, we summarize its features and position. Table 1 compares *VoLearn* with previously developed motion-learning systems. Generally, the motion learning systems can be designed to focus on different motion categories to benefit the different applications. In this paper, we mainly focused on the typical application of a motion-learning system (i.e., rehabilitation and daily exercises). Existing systems can be assessed from two basic aspects.

a) Daily/professional device utilization: Professional devices, such as a MoCap suit, provide the system with reliable input and processing capability while decreasing the overall "ubiquity" in terms of possible implementation.

Table 1. Comparison of our *VoLearn* system with previous motion-learning systems

System	Device	Application	Daily device	Customized motion
NurseCare [18]	Smartphone	Transfer the patient	✓	
Physio@Home [57]	Motion Tracker	Physiotherapy Exercises		
Ayoade et al. [3]	IMU	Knee rehabilitation		
Schönauer et al. [47]	Motion Tracker	General exercises		
YouMove [1]	Kinect	General exercises		✓
Buttussi et al. [5]	Smartphone	Fitness activities	✓	
Pose Trainer [11]	RGB Camera	Posture	✓	
Tutu [56]	Kinect	Ballet dance		
SleeveAR [54]	Motion Tracker	Arm rehabilitation		✓
Onebody [25]	Kinect	General exercises		
My Tai-Chi [23]	HMD	Tai Chi Chuan		
LightGuide [53]	Kinect	Hand motion		
Just follow me [69]	HMD	General exercises		
MotionMA [63]	Kinect	General exercises		✓
COPD Trainer [55]	Smartphone	COPD patients	✓	
VoLearn	**Smartphone**	**General exercises**	✓	✓

In addition, more common-use personal daily devices, such as smartphones and smartwatches, can be used as input or output devices. Although the fidelity of these devices is not as high, daily devices can provide a system with a better fit between a user's daily demand and ease of utilization.

b) Customized/designated motion input: The input motion determines the learned target. Some systems have been designed for specific cases, such as specific exercises for patients [3, 5, 55]. The user cannot interact with the desired motion to obtain a more customized application. In contrast, some other systems support the user interface to help users access the motion and determine a more flexible desired motion for various scenarios [1, 13, 63]. These systems can realize a customized motion for learning.

As shown in Table 1, various systems are combined with professional devices to detect movement and are targeted only at designated input motions. Thus, we position our system as one that uses a daily device and supports customized motions. Specifically, conventional motion-learning systems consider more or less only specific parts of the motion learning process. To benefit the whole learning process, *VoLearn* is a more holistic system that takes into account both the trainer and the trainee and makes contributions on these two ends as well. For the trainer-end, the traditional motion input design normally supports limited editing capability, which restricts the adaptability for novice learners with different physical conditions. However, relying on the 3D motion conversion, we designed an innovative motion editing interface that allows secondary development based on the input motion video. It facilitated the design of a motion which can be customized for various individuals so that the flexibility of the learned object/motion is dramatically increased. While at the trainee end, we also presented a novel feedback design with smartphone-based components. Compared with the traditional feedback design, *VoLearn* using the audio as the metaphorical message helps the user learn a motion with a more interactive understanding and allows a broader range of application scenarios.

4 SYSTEM OVERVIEW

Figure 2 presents an overview of *VoLearn*, whose main design characteristics can be summarized as follows:

a) Operable and interactive motion design: The desired motion can be converted from a 2D RGB video to a 3D avatar. The interface in the virtual 3D environment facilitates the control of the motion's amplitude and speed to help design a customized motion for learning. In addition, the system can generate the output of the motion's data and sensor placement for real-world motion learning.

b) High "ubiquity" feature: As the main feedback is designed relying on a smartphone, the user is able to utilize the auditory feedback as a supporting tool. Furthermore, the auditory feedback removes the necessity of an active area and can be conveniently utilized on more occasions for training exercises.

c) Comprehensive feedback mechanism: An error-based feedback system is developed to help improve a user's learning process. The error between the desired motion's amplitude and speed and those of a novice user is mapped into the audio elements. The feedback aims to reduce errors and helps the user to conduct an accurate motion.

Sections 5–7 describe the three *VoLearn* components in detail: motion editing interface, generation of reference data, and auditory feedback.

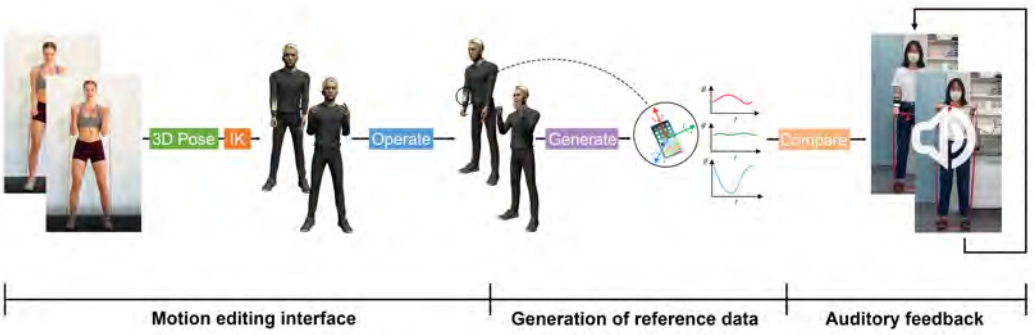

Fig. 2. *VoLearn* begins with a 2D RGB video. Then, the subject in the video can be converted to a 3D virtual humanoid avatar. The 3D avatar can be manipulated by the user to adjust the motion characteristics, and the virtual sensor attached to the 3D body part can generate rotation data. Compared to the real sensor data generated from user motion, it provides continuous auditory feedback. When the user performs a correct motion, an audio prompt is generated.

5 MOTION EDITING INTERFACE

5.1 Virtual 3D Motion

To design a motion, it is important to obtain the key joint position of the body in the 3D environment. We used a state-of-the-art commercially available 3D motion model, *DeepMotion*, to convert the 2D video to a 3D motion file [16]. As the converted file is in "fbx" format, we developed the script to extract the joint position data of the avatar. The central hip (pelvis) was selected as the origin coordinate of the local reference frame. We cut off each motion period from the repeated motion video, and then calculated the average of several motion repetitions to obtain the ultimate motion period. Subsequently, a Kalman filter was used for each joint datapoint to eliminate noise. We then applied the joint position information to the avatar through inverse kinematics to facilitate the rotation of a model mesh with the virtual avatar (Figure 3).

Fig. 3. Illustration of converting the video into animation with different humanoid models (e.g., squatting and Tai-Chi Chuan).

5.2 Interface for Motion Editing

We designed a concise interface for users to create a 3D motion for learning. This system was based on the Unity3D game engine. Motion functionality can be divided into three stages: (a) importing the scene, (b) motion preview and managing the scene, and (c) motion editing and creating the scene.

In the first scene, the user inputs the desired video into the system and names the motion (Figure 4). After processing the vision model and filter unit, the animated motion is shown to the user for preview (Figure 5) and can be stored in the user's local library.

Fig. 4. Interface for processing the input RGB video with custom naming: (a) upload the original video file, (b) video presentation, and (c) video/motion registration with a user-given name.

The function of changing the motion amplitude is then designed. The motion amplitude can be adjusted on the three independent axes. We applied nonlinear mapping to facilitate zooming in and out of the amplitude of the initial imported motion. The value of the *tanh*( ) function was multiplied by the original joint position at the three axes; then, the changed position data were moved to the same start point as that in the initial position series. As we applied nonlinear amplification, there was no over-amplification or reduction in the 3D motion. In addition, slider bars were created to modify the upper-body (arm) and lower-body (leg) amplitudes and speed (Figure 6).

We also designed an interface to combine various motions. In pre-processing, we separated the upper-body motion from the lower-body motion (to save and manage the motions separately). Thus, upper-body motion can

Fig. 5. User preview scene for adding a new motion and managing the local library. Multiple perspectives are also provided for checking the transferred animation: (a) choosing the converted 3D animation from the library, (b) dragging a blank space to rotate the avatar to obtain multiple views, and (c) adding several motions of interest by the user for subsequent operations.

Fig. 6. Motion amplitude adjustment (upper arm used for illustration). Select the desired axis and drag the button of the slider bar of the "Arm Amplitude" to the right. This will increase the range of motion of the arm region. Dragging to the left reduces the motion.

be combined with the lower-body motion to produce a new whole-body motion according to the user's needs [19] (Figure 7). The user can determine the motions that the upper and lower limbs will perform, as well as their frequency, in one period (Figure 8).

6 GENERATION OF REFERENCE DATA

6.1 Virtual Motion Data

Once the desired motion was designed, the virtual sensor's data were output to allow the user to practice with feedback. To capture the moving status of users, we used inertial data. Generally, rotation data or displacement information obtained from acceleration are considered. However, the errors and noise generated from video conversions likely accumulate through the integral operations between acceleration and displacement. Therefore, we selected rotation data. As avatar's joint position data was related to the local reference frame of the pelvis, there was no need to convert the rotation data from the local frame to a global frame.

Because the body hierarchy structure was used, a virtual sensor was attached to the parent node of the body hierarchy skeleton (Figure 9) and followed the movement of the body segment. We exploited the rotation matrix to transform the global sensor Euler angle data to local frame angle data, as presented in Equations (1), (2), and

Fig. 7. Illustration of the motion combination. For example, press the button to enable upper-body movement (chest expansion) to be combined with lower-body movement (squatting).

Fig. 8. Motion combination function. A single upper-body motion (raise hand) can be added to a single lower-body motion (squatting). Pressing the plus button will increase the frequency of the upper-limb motion (e.g., to three). This means that the lower limbs perform one movement, while the upper limbs perform three movements.

(3).

$$R_s(t) = R_{so}^T \cdot R_g(t) \tag{1}$$

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \tag{2}$$

$$\theta_x = \arctan(r_{32}, r_{33}), \theta_y = \arctan\left(-r_{31}, \sqrt{r_{32}^2 + r_{33}^2}\right), \theta z = \arctan(r_{21}, r_{11}) \tag{3}$$

where the R_{so} is the transpose rotation matrix of the initial sensor Euler angle data. $R_g(t)$ is the global rotation matrix at each time. $R_s(t)$ is the rotation matrix transformed from the global frame to the sensor's local frame, and θ_x, θ_y and θ_z are the rotation angle data at the x-,y-, and z-axes in the sensor's local frame, respectively.

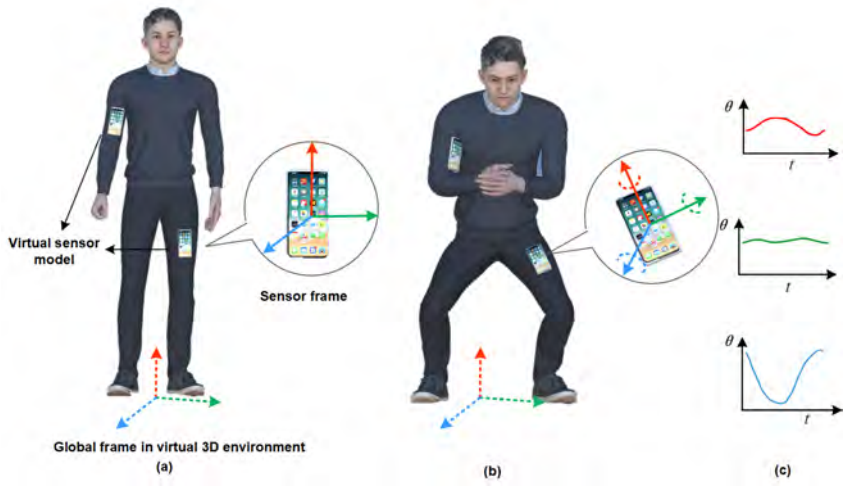

Fig. 9. Illustration of virtual rotation data generation. (a) Two virtual sensor models are attached to the virtual avatar. The sensor's local and global frames are presented. (b) The virtual sensor model is moved according to the attached body segments. (c) The rotation data of the sensor's local reference frame in a complete motion period are output.

6.2 Generating Sensor Placement and Data

We mainly assumed that our system would be applied on a smartphone and that an individual normally keeps one smartphone. Thus, it was necessary to analyze the whole-body motion data and determine which position is the most significant and should be concentrated on first. For this purpose, we developed an algorithm to analyze the rotation data from all possible sensor placements in one defined motion period and generated the recommended placement (as shown in Algorithm 1). The following are the relevant definitions:

a) Effective sensor (smartphone) placement: P_i is the *waist*, *right upper arm*, *right forearm*, *left upper arm*, *left forearm*, *right upper leg*, *right lower leg*, *left upper leg*, and *left lower leg*.

b) Body segment exercise intensity, E_i: The exercise intensity is related to the movement (rotation) velocity of the indicated body segment that is relative to the parent body segment. For example, the intensity of the left knee is the movement (rotation) velocity related to the left upper leg. Each body segment i has three Euler rotation data points, $\theta_i = \{\theta_x, \theta_y, \theta_z\}$. The E_i is defined as a fusion of three-axis rotation velocities in one motion period (i.e., $E_i = \sum_1^n \sqrt{(\theta_{x(k)} - \theta_{x(k-1)})^2 + (\theta_{y(k)} - \theta_{y(k-1)})^2 + (\theta_{z(k)} - \theta_{z(k-1)})^2}$, where n represents one motion frame length, and k indicates the current frame number). The definition of a parent-to-child segment follows the generally used hierarchy structure description of the human body [63].

c) Virtual motion reference data : We defined the data used for comparison with real rotation data as reference data. The reference data are expressed on a single axis, which is defined as the axis of effect (as in [34]), indicating the axis on which the effect of the action is mainly manifested.

6.3 Mapping the Virtual Sensor Data to Real-world Data

After the virtual reference data has been generated, it is essential to map the virtual sensor's data to the real-world data. Three main components output from the virtual sensing environment must be mapped. First, an appropriate wearing position can be indicated by a virtual smartphone's presentation. After the suggested wearing position

Algorithm 1 Sensor placement and reference data generation

Input: $\{P_i\}$, Motion sequence M$\{0, 1, ..., n-1\}$
Output: P_s as the placement, $\{\theta_k\}$ as the reference data

$k \leftarrow 1$
$i \leftarrow 1$
for each $E_i \leftarrow 0$
for each $\theta_{ix(0)} \leftarrow 0, \theta_{iy(0)} \leftarrow 0, \theta_{iz(0)} \leftarrow 0$
while k <n+1 **do**
    for i <10 **do**
        $e_i \leftarrow \sqrt{\left(\theta_{ix(k)} - \theta_{ix(k-1)}\right)^2 + \left(\theta_{iy(k)} - \theta_{iy(k-1)}\right)^2 + \left(\theta_{iz(k)} - \theta_{iz(k-1)}\right)^2}$
        $E_i \leftarrow E_i + e_i$
    end for
end while
$E_{max} \leftarrow max\,\{E_i\}$
$s \leftarrow index\ of\ E_{max}\ in\ \{E_i\}\ as\ the\ suggested\ placement\ number$
$k \leftarrow 1$
while k <n+1 **do**
    $\{x_k\} \leftarrow \theta_{sx(k)}$
    $\{y_k\} \leftarrow \theta_{sy(k)}$
    $\{z_k\} \leftarrow \theta_{sz(k)}$
end while
$\{\theta_k\} \leftarrow max\ Variance\ in\ \{x_k\}, \{y_k\}, \{z_k\}$

is determined, the virtual smartphone model can be presented on the avatar's body segment, and the user can wear the real smartphone according to the presentation.

Following Equation (1), we converted the device's Euler angles from the real global frame to the sensor's local frame with the same installation attitude as that in the virtual avatar. In addition to installation placement, a transformation matrix is also required to ensure that the definition of a virtual sensor coordinate is the same as that of a real sensor's coordinate. For example, when the turning the sensor at the coronal plane, the value of the y-axis is changed, both in the virtual- and real-sensor local frames. The detailed processing is presented in Figure 10.

Finally, the time scale must be made uniform. Because the input video determines the virtual motion's frame rate, we converted the virtual motion's frame rate into a uniform rate to map it to the real sensor. Thus, the virtual sensor had the same sampling frequency as the real sensor. The exploited method is based on a *cubic spline* [70].

7 AUDITORY FEEDBACK

7.1 Real Sensor Development

We developed an application on a personal smartphone to detect rotation information for motion learning. Following a series of easy steps, the user performed the required calibration before capturing the real motion data. The real rotation data were recorded within the smartphone's local frame.

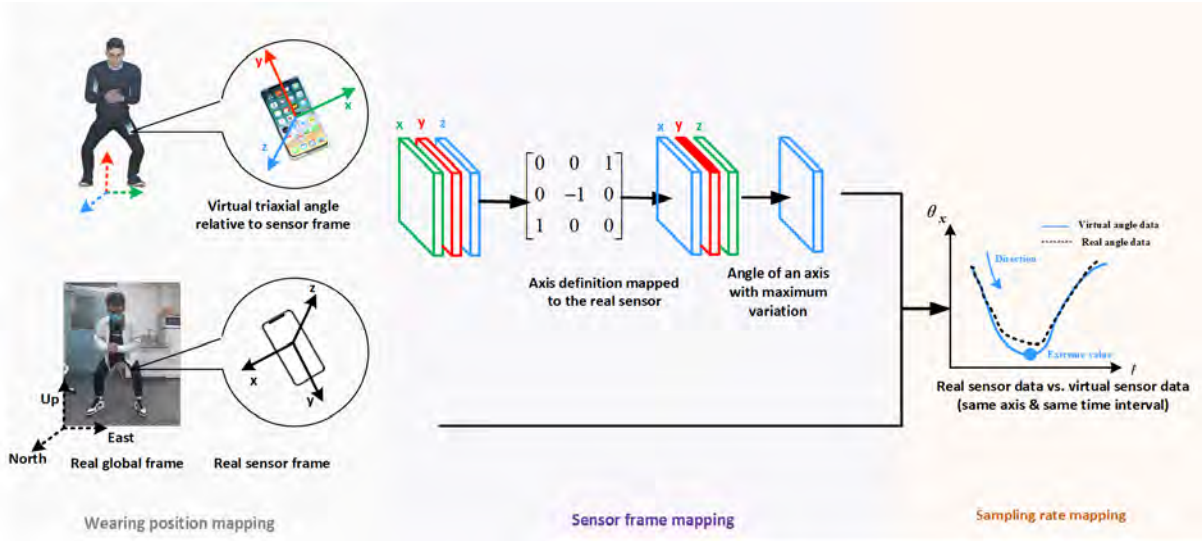

Fig. 10. Domain mapping from the virtual to the real world. The rotation data are first transformed into the virtual sensor's local frame. The three axes of the virtual sensor are then transformed against the real sensor's definition. The axis corresponding to the largest change data is selected as the compared axis/axis of effect. The real rotation data are also transformed from a global frame into a real-sensor local frame.

7.2 Error-based Feedback and Audio Elements

We designed an auditory feedback system to aid in the learning/training process and realize wider applied scenes. Generally, no strict evidence shows which type of feedback modality has absolute advantages [51]. As each individual can determine the learning process, the feedback design should decrease the subjective effect that makes the information as easy to understand as possible and retains interactivity to improve the user's motivation. Therefore, we mainly adopted an audio pitch and cue as the metaphor to realize the motion instruction/feedforward process and verbal information to perform motion assessment. Three types of audio reminders were developed:

a) Pitch variation: When the user starts to perform the motion and approaches the correct position, the audio pitch increases.

b) Cue: Once the user reaches the correct position, an audio cue is generated.

c) Voice: After the user completes one period of motion and returns to the start status, the system provides an assessment and next improvement according to the current motion data.

Figure 11 presents an illustration of instructing the user to conduct the correct motion with different audio elements. The audio pitch and cue can be used to remind the user of the current motion amplitude in real-time; the motion speed/time length is assessed after the whole motion is completed. In addition, in considering both the amplitude and time of motion, the dynamic time warp (DTW) distance between the real motion data and the reference data can be calculated to provide holistic feedback as an assessment of the current motion. The main design principles are shown in Table 2 and introduced as follows:

a) Obtain an entire motion period. Similar to offline analyses [34, 55], we used a tiny sliding window to determine the motion start and endpoints. The length of the window is indicated as l. The variance of the data in a whole window is calculated at each time. When the variance of the total data in the window exceeds the threshold θ_o, it

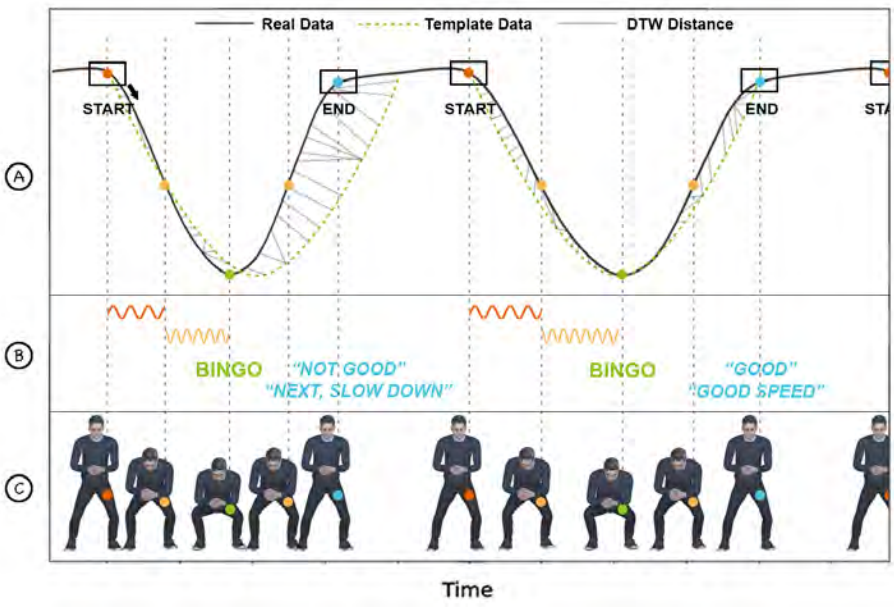

Fig. 11. Demonstration of auditory feedback. (a) Motion data detection. (b) Audio elements. (c) Real motion. When the user starts to perform the motion, the system gives a basic continuous sound (such as "Di Di"). When the user is approaching the correct position, the pitch of the continuous sound increases. Once it reaches the correct position, the system gives a "DingDong"-like sound. After that, the sound disappears. Until one motion period is completed, the auditory feedback will give a holistic assessment and speed improvement.

is recognized as the start point and the current window's mean value m_s is recorded. In contrast, when the next window's variance is less than the threshold and the mean value drops within the range of the initial mean value [m_s-σ_s, m_s+σ_s], the current timestamp is considered the endpoint of the motion. Subsequently, the entire real motion period data are obtained.

b) Calculate the motion time length. After one motion period data have been recorded, the time length of the motion, T_r, can be simply calculated and assessed. As shown in Table 2, λ is the judgment margin of the time length comparison between the real motion and reference motion time, T_m.

c) Calculate the DTW distance. We finally calculated the DTW distance between the reference data and real motion data: $d = DTW\left(\{\theta_m, t_m\}, \{\theta_r, t_r\}\right)$. The calculation result can be mapped into a qualitative assessment through a piecewise function.

8 EVALUATION AND EXPERIMENT

This section introduces the evaluation of the *VoLearn* system, which can be divided into three parts. We first assessed the error between the input motion data and output virtual rotation data. A user study was then conducted to test the performance of the 3D avatar and auditory feedback design. Finally, we evaluated the proposed system with people in motion-related professions to better position the entire system design.

Table 2. Trigger conditions of the audio elements. A flexible margin should be set for each trigger condition, such as ε for target reaching judgment, σ_s for endpoint judgment, and λ for time length judgment.

Object	Trigger condition	Action
Start point	Var $\{\theta_1, \ldots, \theta_l\} > \theta_o$	Start recording real motion data & and obtain the mean value of start window m_s
Pitch	$\theta_i \rightarrow \theta_{target}$	Pitch increase
Cue	$\theta_i \in \left[\theta_{\text{target}} - \varepsilon, \theta_{\text{target}} + \varepsilon\right]$	Generate the audio cue
Endpoint	Var $\{\theta_1, ..., \theta_l\} < \theta_o$ and mean $\{\theta_1, \ldots, \theta_l\} \in [m_s - \sigma_s, m_s + \sigma_s]$	End recording real motion data
Time length assessment	$T_r \leq \lambda T_m$ $\lambda T_m < T_r \leq (1+\lambda) T_m$ $T_r > (1+\lambda) T_m$	"*Next, slow down*" "*Good speed*" "*Next, hurry up*"
Holistic assessment	$d <$ boundary1 boundary 1 $< d <$ boundary 2 $d >$ boundary2	"*Excellent*" "*Good*" "*Not good*"

8.1 Study 1: End-to-end Test of Signal Error on Output Body Segment

Because *VoLearn* is an end-to-end system that converts a 2D video to an audio output, we first evaluated the error between the input body segment rotation data and output virtual body segment rotation data, to test how much raw error the system has when not in use. Ten motions were selected from the typical motion tutorial from YouTube, which involved typical aerobic motion, yoga, and anaerobic exercises. One developer wore nine smartphones (Huawei Mate 10) on their bodies, with our developed application operating on all smartphones. While conducting the motion, the tester was simultaneously recorded by a web camera. The involved motions are presented in Figure 12.

After the videos were recorded, we processed them using our system and output the virtual sensor data with the designated body segment. Then, we calculated the root mean square error (RMSE) between the related input body segment's rotation data and output virtual sensor data in the same motion period (100 frames). We also produced an extreme value error between the real and virtual data, which determined the motion amplitude.

Table 3 presents the results of the examined motions. From input to output, the errors likely originated from the 3D motion conversion by the vision model, filtering to smooth the motion, and wearing deviations between the real and virtual smartphones. Overall, the motions that contained body segment overlapping and partial blocking (like motion 10) had a relatively larger RMSE over the entire period. In contrast, most of the motions retained the lower amplitude error. As a designing tool, our system's motion modification function can adjust the amplitude of movement and, thus, compensate for the initial error to some extent. Thus, as long as the generated 3D motion does not destroy the original motion much, the system can be used for our purposes.

8.2 Study 2: User Study of 3D Avatar and Auditory Feedback Design

VoLearn was designed using two main factors: the 3D avatar and auditory feedback. Both these factors can affect a user's behavior and study condition. Thus, in this section, we describe a user study conducted to not only assess the learning result but also to evaluate the two factors related to the learning process, especially the effect of auditory feedback.

Fig. 12. Tested ten motions.

Table 3. Test result of 10 motions of output virtual rotation data compared to smartphone rotation data related to the input motions.

Motion	1	2	3	4	5	
Suggested Sensor Position	Left upper arm	Left upper leg	Left upper leg	Waist	Right upper leg	
RMSE (degree)	14.93	9.62	19.29	5.61	15.52	
Amplitude Error (degree)	0.08	4.07	12.76	6.04	1.35	
Motion	6	7	8	9	10	**Average**
Suggested Sensor Position	Left upper leg	Left upper leg	Left upper leg	Left upper leg	Right forearm	/
RMSE (degree)	3.79	15.48	14.02	16.58	56.15	**17.01**
Amplitude Error (degree)	2.91	12.27	5.33	8.81	17.89	**7.18**

8.2.1 Task. This study was designed to complete a motion-learning process. Therefore, considering the human and timing resources, we took six typical motions (motions 1–6) from Study 1 with various categories and gave them to the participants to study the motions.

8.2.2 Hypotheses. As introduced in section 4, the auditory feedback interface was implemented on a smartphone; thus, we assumed two usage scenarios for this system. (a) is the main applied scenario of *VoLearn*. The users learn the motion with *VoLearn* (by the 3D avatar and auditory feedback) and then do the motion with auditory feedback (no 3D avatar but with the smartphone). Such occasions might occur at a gym, home, or exercise training location, where the user desires guidance. (b) is a wider applied context. The users learn the motion with *VoLearn* (by the 3D avatar and auditory feedback) and perform the motion without auditory feedback (no 3D avatar and no smartphone, i.e., device-free). This occasion can occur anywhere the user wants to conduct a motion but does not

need guidance. Accordingly, to examine the system's efficacy in these two scenarios, we posited the following hypotheses.

H1a: The *VoLearn* system is helpful in reducing the error of motion amplitude and time in the scenario (a).

H1b: The *VoLearn* system is helpful in reducing the error of motion amplitude and time in the scenario (b).

As the completion of the motion-learning process requires several phases [50], the motion complexity can affect the key factors in the learning process, such as attention and body control ability. In addition, users might have different learning abilities under different given motions. However, there is generally no strict definition of a motion's complexity. Thus, we considered that multi-body movement is more complex than single/symmetrical limb movements because the user needs to concentrate on understanding the motion and coordinating the body segment [62]. The tested motions are divided into simple and complex motions, as shown in Figure 13. For simple motions, the users were expected to understand the motion and control their bodies easily. For complex motions, as multiple body segments require movement, the users were expected to find it difficult to control the motion amplitude and time accordingly. Therefore, to better figure out the system's effect, we examined the system with two types of motions in scenario (a), which highlights the main scenario of *VoLearn*. We examined the system to address the following hypothesis:

H2: The *VoLearn* system is more effective in learning complex motions than simple motions in its main scenario.

8.2.3 Participants. We recruited 20 participants (4 female and 16 male) with an average age of 23.8 (SD = 1.73) from different organizations to join the experiment. Each participant was paid 10 dollars. They were young adults who did not have daily exercise regimes.

8.2.4 Design. We designed a mixed experiment with *auditory feedback* and *motion complexity* as independent variables. Two main stages were involved in the experiment, the train stage and test stage, which corresponded to scenarios (a) and (b) respectively. Participants studied the motion in the train stage and presented the learning outcome in the test stage. The study protocol was presented in Figure 13. The participants who did not use auditory feedback were classified as Group NN (no auditory feedback in the train stage and no auditory feedback in the test stage). To examine H1a and H1b, people using auditory feedback formed two sub-groups related to two different scenarios, i.e., Group AA that corresponds to the scenario (a) (use auditory feedback in the train stage the motion and use auditory feedback in the test stage) and Group AN that corresponds to the scenario (b) (use auditory feedback in the train stage and no auditory feedback in test stage. Thus, Group NN had 10 testers (8 male and 2 female with an average age of 24.2 (SD = 1.8)) and both Group AA and Group AN also had 10 testers each (the same participants; 8 male and 2 female with an average age of 23.6 (SD = 1.6)). The formulation of these groups is also presented in Figure 13.

8.2.5 Measurement. a) *Motion amplitude and time error.* As the feedback provided the motion amplitude and speed information, two main metrics were recorded: the motion amplitude error and time error compared to output motion data from the virtual sensor. It determines how close the user's motion is in time and space to the given motion. To obtain these data, we attached the smartphone to each participant's designated body segment (calculated by the system itself) and requested them to conduct the motion three times to record the rotation data and time data from *VoLearn* on the smartphone. We averaged the amplitude error and time error from the three recordings of each motion. If auditory feedback helps the user in motion learning, the amplitude and time error should be smaller compared with the group that does not use auditory feedback. In particular, if there is a significant error reduction between Groups NN and AA, then H1a is supported. Similarly, if there is a significant error reduction between Groups AN and NN, then H1b is supported. Considering the auditory feedback and motion complexity, if the significant error reduction between the Groups AA and NN in complex motion learning is larger than significant error reduction between the Groups AA and NN in simple motion learning, the system is considered to be more effective for complex motion learning, thus supporting H2.

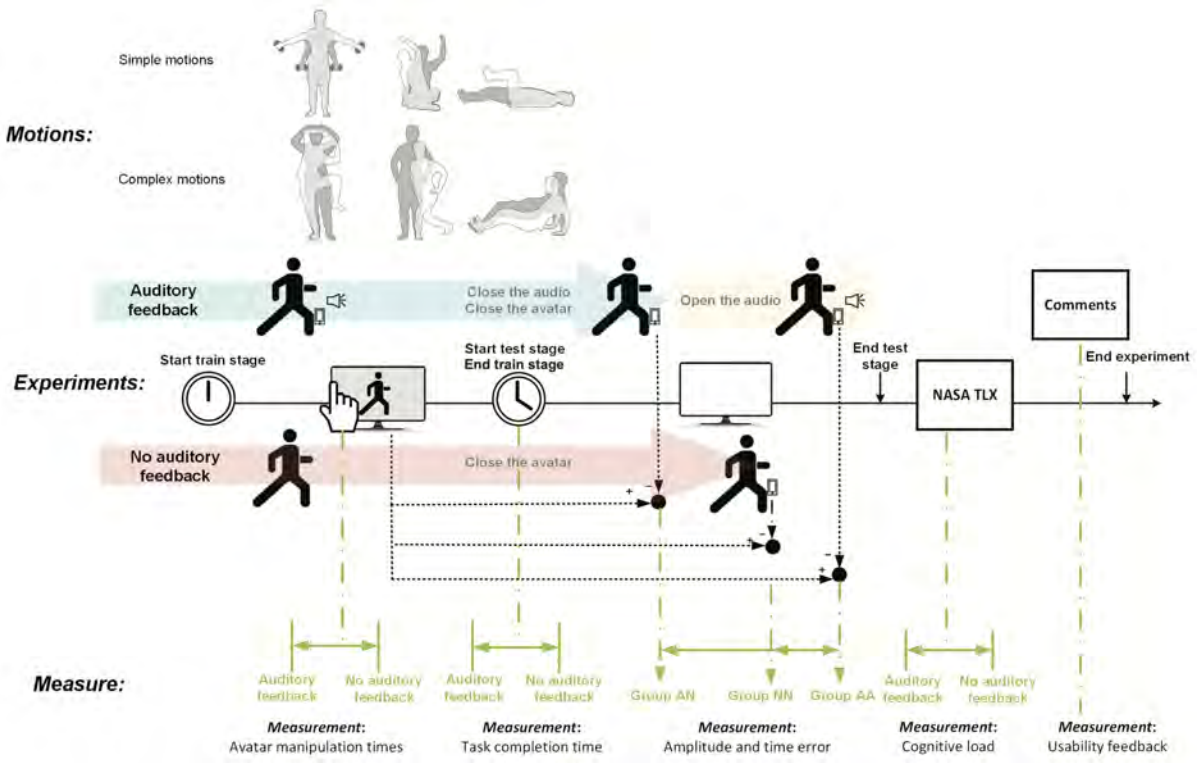

Fig. 13. Illustration of user study design. First, the participants are divided into two groups: one uses auditory feedback when learning the motion and the other does not use it (Group NN). Two tests are conducted in the test stage for the group using auditory feedback, forming two sub-groups, AA and AN.

b) *Cognitive load.* The cognitive load was measured to assess whether the additional auditory feedback increased the user's attention and mental burden during the train stage. We used the NASA-TLX questionnaire [24] to measure cognitive load and compared the participants who used and did not use auditory feedback.

c) *Avatar manipulation times.* A 3D avatar providing the visual information helps the user become familiar with the desired motion. Unlike traditional 2D video sources, the 3D avatar can present multi-perspective information regarding the motion. Thus, we measured the number of times the user manipulated the avatar. In addition, we compared these times between the participants with and without auditory feedback. This can evaluate whether auditory feedback decreases the requirement of visual information for the user in the train stage.

d) *Task completion time.* As the feedback aimed to reduce the motion's amplitude and time error, we did not limit the time for users to study the desired motion. We recorded the time for each individual after we showed the 3D avatar and ended the recording when the participant self-reported the end of the motion study.

e) *User comments.* To better determine the usability of auditory feedback, we asked the participants to comment on our system's usability.

8.2.6 Procedure. Each participant was requested to complete a consent form and a demographic investigation. We first introduced the 3D avatar and the operation to change the perspective. For the participant using auditory

feedback (Group AA/AN), we introduced the feedback component, and each participant was then familiarized with the system through a case study (in addition to task motions). After the participant reported that they had understood the system, we began the experiment. We asked the participants to wear our experimental smartphone (Huawei Mate 10) in the corresponding position output requested by the system. After calibration, the participant began the motion learning from the virtual 3D avatar and the reception of auditory feedback. For Group NN, the participant started motion learning from virtual 3D avatar only.

When a participant subjectively reported that they completed the learning, we recorded the total study time and the number of times the participants changed their view of the avatar. Then, we closed the 3D avatar and the participants entered the test stage. For Group NN, we requested that the participant conduct the motion three times and recorded the motion data using the worn smartphone. For the participants who used auditory feedback, we first requested them to perform the motion three times without audio (from Group AN) and then turn on the audio and perform the motion three more times (from Group AA). To avoid the additional training effect, Group AN was tested before Group AA.

After finishing the final result recording, all group participants were requested to answer the NASA TLX questionnaire. Finally, we asked Group AA/AN participants to comment on the system. The entire experimental process was recorded using a camera for archival purposes.

8.2.7 Methodology. We adopted both qualitative and quantitative methods to analyze the results. Because the variables did not present the characteristics of normal distributions, to examine our hypotheses, we conducted the Aligned Rank Transform (ART) ANOVA test with *auditory feedback* and *motion complexity* as two independent variables. In addition, the cognitive load and task completion time were compared with the Mann-Whitney U test to compare Groups AA/AN and NN.

8.2.8 Result. a) *Results for H1a and H1b.* The result of amplitude and time error for three groups was shown in Figure 14. Regarding the amplitude error, the ART ANOVA showed a significant main effect of *auditory feedback* ($F_{2,33} = 4.16$, $p < 0.05$) while no significant effect of *motion complexity* and no significant interaction effects between these two variables. A post-hoc test with Holm correction of *auditory feedback* main effect showed the significant differences between Groups AA and NN ($p < 0.05$) and Groups AN and NN ($p < 0.05$). Moreover, there was also only a significant main effect of *auditory feedback* ($F_{2,37} = 18.65$, $p < 0.001$) in terms of the time error. The post-hoc test showed significant differences between Groups AA and NN ($p < 0.001$) and between Groups AN and NN ($p < 0.001$). These results validated hypotheses H1a and H1b. The auditory feedback successfully helped users to learn the motions, controlling their body movement amplitude and speed.

b) *Results for H2.* Considering both the *auditory feedback* and *motion complexity*, there were no significant interaction effects between these two variables. Thus, the post-hoc test was not conducted for pairwise checking on the specific groups. So, the H2 was not supported. For the motion complexity, as we mentioned in 8.2.2, it is hard to define it due to individual differences. Thus, in the experiment, we have roughly classified complex and simple motions in terms of intensity of body movement, selecting typical body motions. Also, since we arbitrarily chose the motions, there was a chance that some of them were too hard to perform for some participants. Therefore, the rejection of H2 was only for the complex and simple motions covered in this paper. We conjecture that further studies based on more sophisticated and representative definition of motion complexity could reveal the efficacy of *VoLearn*.

c) *Cognitive load, avatar manipulation times, and task completion time.* Figure 15 reports the results obtained from NASA TLX, the manipulation times of the avatar, and the task completion time. With the Mann-Whitney U test, only the task completion time revealed a significant difference. That is, the participants who used auditory feedback spent more time on the learning. This is because the participants normally took time to improve their movement according to the feedback. We aimed to correct the user movement and perform a more accurate motion when learning. Thus, we can reasonably assume that the participants will consume more time to achieve a

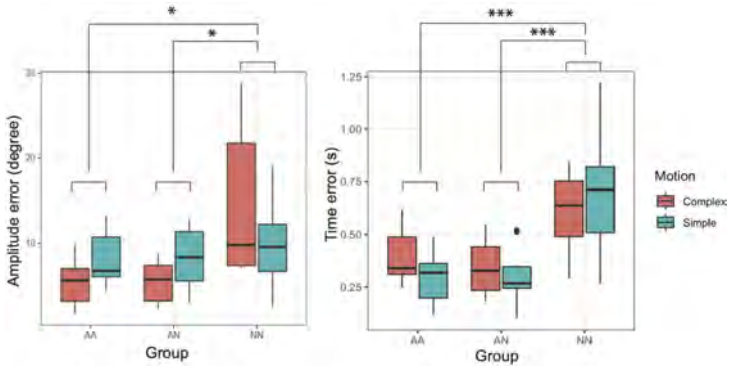

Fig. 14. Results of amplitude and time errors among the three groups.

better learning outcome due to the introduced feedback. Note that there was no significant difference in cognitive load, which means that *VoLearn* does not impose additional load even if users spend more time leveraging auditory feedback.

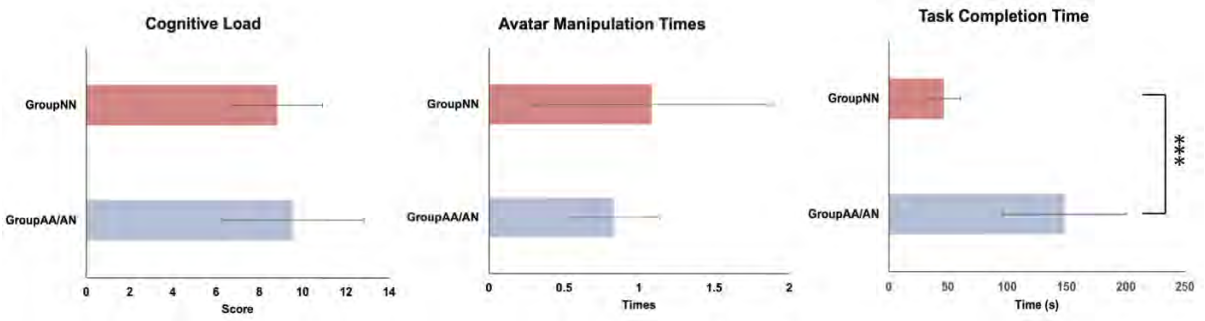

Fig. 15. Qualitative evaluation results. We find a significant difference in the task completion time. The error bar indicates the standard deviation ***: p<0.001.

d) *User comments.* As the result supported H1a, H1b, the system was validated to decrease the error during the learning with the auditory feedback used. Here, we report the obtained comments about usability.

All participants were positive about the system and recognized that it could help them perform a motion better. Some user feedback addressed the general value of the system, such as "*The system can help in understanding the motion.*" and "*The system gives me a goal to train the motion and makes me feel not such boring.*" Regarding the audio elements, we used a voting question to collect the participants' favorites. The verbal feedbacks providing the improvement information (such as "*Next, hurry up*") and holistic assessment (such as *"Good/Not Good"*) were tied as the most popular element, and the audio cue followed them. In addition, no one selected the audio pitch. The participants release a lot like about the verbal feedbacks of improvement and audio cue, stating that "*The improvement feedback encourages me to reach a higher standard and motivates me.*" and "*The audio cue can remind me to understand the motion.*"

In addition, we asked the participants about their feelings when using our avatar and auditory feedback compared to watching a conventional tutorial video in general. Most of the participants expressed enjoyment of using our system and felt that the system provides considerable interaction. One exception was a participant who held a more neutral viewpoint: "*I would feel this system was more useful if I could obtain additional information for teaching me which muscle is supposed to be used. This point normally appears in a video tutorial and helps me to understand what I am training.*" One participant suggested that the system could have more contextual elements to make a whole movement more understandable.

8.3 Study 3: Evaluation by Professional Users

Study 2 mainly investigated the design from the perspective of the end user. We fixed all conditions so that they did not allow the participants to use the speed, amplitude, and motion combination functions. The main reason for this is that normal users generally have no goals regarding adjusting the target motion. Therefore, in the third study, we evaluated the function of motion operations with three motion-area-related experts through interviews and demonstrations. The experts operated our system through remote desktop control in an online meeting room (ZOOM) to enable them to use every function. One of the developers demonstrated the auditory feedback with the expert's operation. Afterward, we conducted a semi-structured interview to obtain feedback regarding our system

8.3.1 Feedback from a Gym Coach. We first invited a gym coach to use our system. The coach commented, "*I think that the video input function is really useful to me.*" He thought that this function would help him provide personalized guidance to the students. He first found or recorded a motion video by himself and then converted it to a 3D avatar. He felt that the amplitude adjustment was better and more helpful than the speed adjustment. He said, "*The amplitude adjustment helps me to design a better fitting motion for each student" and "The speed seems not very useful in gym exercise. We normally do not need the student to control the speed precisely.*" He then described how he would see using the system as follows:

"I think the system is more meaningful for remote studying. Except for these functions, I would like the system to have more management on exercise training. And the remote studying feature can also be fused into an exercise training plan so that I can remotely monitor the completion status of a student." Regarding the auditory feedback, he held the view that *"generally, it is very hard to train by yourself without others' guidance. The same as your system, it only tracks one body segment, and is hard to get feedback on the full body. But, it can be used as an additional tool for a student after class. During class, I can train the student and give an indication mainly focused on body parts. And after class, the student can do the exercise by himself with effective feedback to ensure a good motion."*

The coach also indicated interest in the motion combination function. However, he hoped that more sensors could be used to monitor both the upper and lower limbs' motions simultaneously.

8.3.2 Feedback from a Clinical Nurse. The second professional user was a medical nurse. She was experienced in the postoperative care of patients and was usually required to help patients perform simple limb motions to help in their recovery. She stated, "*I really love the auditory feedback design. It will help both our nurses and patients in doing a good recovery motion.*" She said that she had previously only studied the standard motion requirements (like how much of an angle the joint should bend) from a reference book and the approximate location of the motion. She reported, "*Actually, before, we did not know whether the patient had done a correct motion or met the standard. All the judgment was based on empirical data.*" Regarding our system, she suggested, "*And also, the auditory feedback can help a patient get rid of the visual need. Because some patients have to lie on the bed, it is hard for them to watch a video for motion guidance.*" She also thought the speed and amplitude adjustment was useful, stating, "*Particularly, the speed adjustment, it will be useful to help some hemiplegia patients to recover their control feasibility after surgery. The amplitude adjustment will help the performance of a personalized recovery*

scheme by different patients." Referring to the motion input, she said, "*I think supporting the video input is not really useful to me. I hope the motion library can have several predefined standard motions, and we only need to design the different speeds and amplitude for each patient.*" She also liked the motion combination function. "*The combination of two separate motions can be helpful for brain surgery patients in training the coordinating ability of different body limbs. Even if currently one sensor cannot track full-body motion, observation and imitation of the avatar is also an alternative to contribute to the coordination training.*"

8.3.3 Feedback from a Rehabilitation Therapist. The last interviewee was a therapist in a rehabilitation hospital. She held views that were similar to those of the second interviewee. She stated, "*I think the motion combination design helps to train the physical coordination.*" She summarized her feelings toward the whole system as follows:

"The system can play a smaller role during our work. Because we normally meet the patients three or four times a week, patients can get correct guidance under our supervision without an additional device. But, I think it can help the patients to do the motion at their homes. We normally have that situation that the patient calls us seeking guidance after they have gone back to their home from the hospital. If they can use it, it will relieve some of our burdens. And under this condition, the adjustment tool expresses a meaning that helps us create a customized motion intensity for each patient for their home training."

In addition, she thought that supporting the video input was not really useful and shared the same reason as that shared by the second interviewee.

9 DISCUSSION AND FUTURE WORK

9.1 Application Scenario and Characteristics

The whole system can be used to effectively control a motion's speed and amplitude in a supported mode. Thus, more customized applications can be realized in the future based on the current *VoLearn* system, such as customized remote exercise teaching, personalized motion management, home-based rehabilitation monitoring, an exercise platform for vision-impaired individuals, and a portable mobile motion guidance system.

Compared to the popular camera-based system, vision detection can ensure complete human body information, while our system currently only supports one-sensor detection. However, considering the feedback, the provided information is still specific to a particular body segment, as a training exercise normally has a targeted body part for users. Additionally, even if the current camera-based systems have been implemented on a smartphone to enable more ubiquitous applications, users still need to find a suitable position to place the phone and be active in the designated area. It is still somewhat inconvenient. Thus, auditory-based *VoLearn* can facilitate the training process to eliminate the specific space requirement and have wider application cases.

Although the main feedback system can be realized on a smartphone, in our user study, we still provided the 3D avatar to the participants in the train stage. This is because when the user is unfamiliar with a motion, they require visual information as the main source to obtain the overall feeling of the motion itself. It is still difficult to convey a new motion without visual data. However, our auditory feedback can be recognized as a support tool during the motion-learning period to help the user better master a motion. Auditory feedback should be used after ensuring that the user already has a certain degree of familiarity with the motion, and a smartphone helps the user train a familiar motion in a more flexible scenario. In the future, to provide a fully phone-dependent system, the motion-editing system can be combined with the Cloud and transferred to the smartphone end to realize the 3D motion conversion and virtual data output. At the trainee end, depending on different situations, the user can watch the motion to become familiar and then use the audio to improve the learning process. Alternatively, they can use the system combined with the monitor to watch the avatar while simultaneously using auditory feedback.

9.2 Improving 3D Motion Reconstruction

Our system relies on transforming a video stream into 3D animation frames. Because of the leveraged vision model, the performance of the derived 3D motion is further affected by the utilized network. However, as the motion modification tool can perform 3D-direction modification, it can also be used to compensate for the error caused by 3D conversion. Because our design is a type of application of 3D motion extraction from a video, the vision model has not been our focus so far. The system is supposed to ensure that the 3D vision model works as a premise and obtain a reliable motion source (such as selecting a clear video and less body blocking). Thus, we conducted all evaluations under a reliable 3D conversion. We proposed the current pipeline to enable a more ubiquitous and customized motion-learning process. With the activation of more techniques related to the obtained 3D motion (e.g., high-performance 3D motion conversion from a 2D video and convenient RGB-D motion recording), we believe that more streamlined applications can be realized.

9.3 Enabling Further Motion Edition

Currently, we have designed motion amplitude adjustment through three-dimensional directions on the upper or lower body. However, more precise adjustment can be realized by controlling a single joint's 3D position. This means that the user can design a specific limb's motion, which is beneficial to a highly joint-targeted training motion.

Additionally, the motion combination function has attracted much interest from professional users. In contrast, tracking the combined motions would require more markers and information to provide feedback. The next research point would be to develop an efficient and concise feedback to track more body limbs. Furthermore, we envision that a professional motion library can be built from various motions. Each individual can access the library and design combination/personalized target motions for their students.

9.4 Feedback Design

Although the effectiveness of feedback can be demonstrated, the precision of amplitude control is affected by the users' ability to respond to the audio cue. Because an individual is likely to present different reaction abilities to the feedback stimulation, realizing precise amplitude control considering the user's reaction can be the next step.

The current auditory feedback design uses audio as a metaphor, mapping the user's motion into an audio pitch and a cue. The metaphor we used aimed to present an easily understandable and interesting way to help users obtain instruction and improve their motivation. Compared to the most direct way, like using voice to tell the user what to do in the next step, the pitch and cue avoid understanding bias and pauses during movement. It can help to form interactive feedback that does not destroy the integrity of a motion period and facilitates evaluation of the entire motion. Our current design maps the 3D motion into an axis-of-effect movement and provides uniform instruction, which is also helpful in providing an intuitive and general feedback interface. To realize more refined motion learning feedback, it can work on the remaining characteristics of audio to create more information metaphors, such as varying the volume combined with music and using spatial audio with more direction information. However, considering the acceptance degree at the user end, a detailed study is still needed to determine a balance between providing more error-based messages and users' understanding and responsiveness.

9.5 Sensor Position and Number

Considering the practical situation, the current system only supports the indicated body segment feedback because an individual carries only one smartphone normally. So, a sensor placement suggestion was implemented. However, as individuals can present various physical conditions and exercise manners, we did not validate whether the one-sensor placement is the best placement to learn the motion most effectively. It is also unconvincing to

provide a sensor position that is most efficient and can fit each individual. Therefore, we followed a more objective analysis approach to determine the body segment with the greatest intensity of movement and recognized it as the most significant part that needs maximum attention.

9.6 Learning Effect

Learning motion and training are generally long-term effects. However, in our system, we did not intend to discuss a long-term study result. As the feedback system has been implemented on smartphone-based audio information, we believe it is more convenient for the trainee to train a motion over a long-term period. Thus, in our user study, we mainly investigated the immediate effect of the trainee's result with and without using our feedback. In addition, studying the long-term learning effect can be helpful to explore an effective approach to facilitate better training.

10 CONCLUSION

This paper designed *VoLearn*, a system that supports cross-modal motion learning from video input to audio output. Our system converts human motion extracted from a tutorial video to a 3D virtual model and enables more interaction with the desired motion; the operability of the virtual motion was realized, such as speed adjustment, amplitude modification, and motion combination. Furthermore, we added auditory feedback at the user end to instruct the trainee to better understand the correct motion. The evaluation demonstrated the effectiveness of our system and indicated the possible application scenarios. With our developed processing flow and functions, our system can help more people in their motion learning, especially in providing a more customized learning environment.

ACKNOWLEDGMENTS

We would like to thank the Takumi Yamamoto for the support in user study. This work was supported by JST PRESTO Grant Number JPMJPR2134.

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*. ACM, New York, NY, USA, 311–320.

[2] Riku Arakawa and Hiromu Yakura. 2021. Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15.

[3] Mobolaji Ayoade and Lynne Baillie. 2014. A novel knee rehabilitation system for the home. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, New York, NY, USA, 2521–2530.

[4] Shiri Azenkot, Cynthia L Bennett, and Richard E Ladner. 2013. DigiTaps: eyes-free number entry on touchscreens with minimal audio feedback. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, New York, NY, USA, 85–90.

[5] Fabio Buttussi, Luca Chittaro, and Daniele Nadalutti. 2006. Bringing mobile guides and fitness activities together: a solution based on an embodied virtual trainer. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*. ACM, New York, NY, USA, 29–36.

[6] Ryan Canales and Sophie Jörg. 2020. Performance Is Not Everything: Audio Feedback Preferred Over Visual Feedback for Grasping Task in Virtual Reality. In *Motion, Interaction and Games*. ACM, New York, NY, USA, 1–6.

[7] Yuanzhi Cao, Xun Qian, Tianyi Wang, Rachel Lee, Ke Huo, and Karthik Ramani. 2020. An Exploratory Study of Augmented Reality Presence for Tutoring Machine Tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13.

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

[9] Filippo Casamassima, Alberto Ferrari, Bojan Milosevic, Laura Rocchi, and Elisabetta Farella. 2013. Wearable audio-feedback system for gait rehabilitation in subjects with Parkinson's disease. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, New York, NY, USA, 275–278.

[10] Yao-Jen Chang, Shu-Fang Chen, and Jun-Da Huang. 2011. A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. *Research in developmental disabilities* 32, 6 (2011), 2566–2570.

[11] Steven Chen and Richard R Yang. 2020. Pose Trainer: correcting exercise posture using pose estimation. *arXiv preprint arXiv:2006.11718* (2020).

[12] Philo Tan Chua, Rebecca Crivella, Bo Daly, Ning Hu, Russ Schaaf, David Ventura, Todd Camill, Jessica Hodgins, and Randy Pausch. 2003. Training for physical tasks in virtual environments: Tai Chi. In *IEEE Virtual Reality, 2003. Proceedings.* IEEE, Long Beach, CA, USA, 87–94.

[13] Christopher Clarke, Doga Cavdir, Patrick Chiu, Laurent Denoue, and Don Kimber. 2020. Reactive Video: Adaptive Video Playback Based on User Motion for Supporting Physical Activity. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 196–208.

[14] Alexandra Covaci, Anne-Hélène Olivier, and Franck Multon. 2015. Visual perspective and feedback guidance for vr free-throw training. *IEEE computer graphics and applications* 35, 5 (2015), 55–65.

[15] Fabian Lorenzo Dayrit, Yuta Nakashima, Tomokazu Sato, and Naokazu Yokoya. 2014. Free-viewpoint AR human-motion reenactment based on a single RGB-D video stream. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Long Beach, CA, USA, 1–6.

[16] DeepMotion. 2021. DeepMotion. https://www.deepmotion.com/.

[17] William Delamare, Thomas Janssoone, Céline Coutrix, and Laurence Nigay. 2016. Designing 3D gesture guidance: visual feedback and feedforward design options. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, New York, NY, USA, 152–159.

[18] Maximilian Dürr, Carla Gröschel, Ulrike Pfeil, and Harald Reiterer. 2020. NurseCare: Design and'In-The-Wild'Evaluation of a Mobile System to Promote the Ergonomic Transfer of Patients. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13.

[19] Hakuei Fujiyama, Mark R Hinder, Matthew W Schmidt, Michael I Garry, and Jeffery J Summers. 2012. Age-related differences in corticospinal excitability and inhibition during coordination of upper and lower limbs. *Neurobiology of aging* 33, 7 (2012), 1484–e1.

[20] Elizabeth Gibbons. 2007. *Teaching dance: The spectrum of styles.* AuthorHouse.

[21] GmbH. 2016. Kaia Health. https://www.kaiahealth.com/virtual-physical-therapy/.

[22] Natsuki Hamanishi and Jun Rekimoto. 2020. PoseAsQuery: Full-Body Interface for Repeated Observation of a Person in a Video with Ambiguous Pose Indexes and Performed Poses. In *Proceedings of the Augmented Humans International Conference*. ACM, New York, NY, USA, 1–11.

[23] Ping-Hsuan Han, Yang-Sheng Chen, Yilun Zhong, Han-Lei Wang, and Yi-Ping Hung. 2017. My Tai-Chi coaches: an augmented-learning tool for practicing Tai-Chi Chuan. In *Proceedings of the 8th Augmented Human International Conference*. ACM, New York, NY, USA, 1–4.

[24] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[25] Thuong N Hoang, Martin Reinoso, Frank Vetere, and Egemen Tanin. 2016. Onebody: remote posture guidance system using first person view in virtual environment. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. ACM, New York, NY, USA, 1–10.

[26] Eve Hoggan, Andrew Crossan, Stephen A Brewster, and Topi Kaaresoja. 2009. Audio or tactile feedback: which modality when?. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, New York, NY, USA, 2253–2256.

[27] Atsuki Ikeda, Dong-Hyun Hwang, and Hideki Koike. 2018. Real-time Visual Feedback for Golf Training Using Virtual Shadow. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*. ACM, New York, NY, USA, 445–448.

[28] Florian Jeanne, Yann Soullard, Ali Oker, and Indira Thouvenin. 2017. EBAGG: Error-based assistance for gesture guidance in virtual environments. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, Long Beach, CA, USA, 472–476.

[29] Florian Jeanne, Indira Thouvenin, and Alban Lenglet. 2017. A study on improving performance in gesture training through visual guidance based on learners' errors. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. ACM, New York, NY, USA, 1–10.

[30] Rushil Khurana, Karan Ahuja, Zac Yu, Jennifer Mankoff, Chris Harrison, and Mayank Goel. 2018. GymCam: Detecting, recognizing and tracking simultaneous exercises in unconstrained scenes. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–17.

[31] Seung-Jae Kim, Marie Aimee Kayitesi, Amy Chan, and Kimberli Graham. 2017. Effects of partial absence of visual feedback information on gait symmetry. *Applied psychophysiology and biofeedback* 42, 2 (2017), 107–115.

[32] Anna Klapetek, Mary Kim Ngo, and Charles Spence. 2012. Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Attention, Perception, & Psychophysics* 74, 6 (2012), 1154–1167.
[33] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. 2019. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Long Beach, CA, USA, 8658–8667.
[34] Yousef Kowsar, Masud Moshtaghi, Eduardo Velloso, Lars Kulik, and Christopher Leckie. 2016. Detecting unseen anomalies in weight training exercises. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*. ACM, New York, NY, USA, 517–526.
[35] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.
[36] Jiann-Der Lee, Chung-Hung Hsieh, and Ting-Yang Lin. 2014. A Kinect-based Tai Chi exercises evaluation system for physical rehabilitation. In *2014 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, Long Beach, CA, USA, 177–178.
[37] Jeff Lieberman and Cynthia Breazeal. 2007. TIKL: Development of a wearable vibrotactile feedback suit for improved human motor learning. *IEEE Transactions on Robotics* 23, 5 (2007), 919–926.
[38] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. When video meets inertial sensors: zero-shot domain adaptation for finger motion analytics with inertial sensors. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. ACM, New York, NY, USA, 182–194.
[39] Jean-Luc Lugrin, David Obremski, Daniel Roth, and Marc Erich Latoschik. 2016. Audio feedback and illusion of virtual body ownership in mixed reality. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*. ACM, New York, NY, USA, 309–310.
[40] Azumi Maekawa, Shota Takahashi, MHD Yamen Saraiji, Sohei Wakisaka, Hiroyasu Iwata, and Masahiko Inami. 2019. Naviarm: Augmenting the Learning of Motor Skills using a Backpack-type Robotic Arm System. In *Proceedings of the 10th Augmented Human International Conference 2019*. ACM, New York, NY, USA, 1–8.
[41] Zoe Marquardt, João Beira, Natalia Em, Isabel Paiva, and Sebastian Kox. 2012. Super Mirror: a kinect interface for ballet dancers. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1619–1624.
[42] Troy L McDaniel, Morris Goldberg, Shantanu Bala, Bijan Fakhri, and Sethuraman Panchanathan. 2012. Vibrotactile feedback of motor performance errors for enhancing motor learning. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, New York, NY, USA, 419–428.
[43] Stephen W Mereu and Rick Kazman. 1997. Audio enhanced 3D interfaces for visually impaired users. *ACM SIGCAPH Computers and the Physically Handicapped* 1, 57 (1997), 10–15.
[44] Peter Mohr, David Mandl, Markus Tatzgern, Eduardo Veas, Dieter Schmalstieg, and Denis Kalkofen. 2017. Retargeting video tutorials showing tools with surface contact to augmented reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 6547–6558.
[45] Cecily Morrison, Neil Smyth, Robert Corish, Kenton O'Hara, and Abigail Sellen. 2014. Collaborating with computer vision systems: An exploration of audio feedback. In *Proceedings of the 2014 conference on Designing interactive systems*. ACM, New York, NY, USA, 229–238.
[46] PumpOne. 2020. FitnessBuilder. https://www.pumpone.com/fitnessbuilder/.
[47] Christian Schönauer, Kenichiro Fukushi, Alex Olwal, Hannes Kaufmann, and Ramesh Raskar. 2012. Multimodal motion guidance: techniques for adaptive and dynamic feedback. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, New York, NY, USA, 133–140.
[48] Yoones A Sekhavat and Mohammad S Namani. 2018. Projection-based ar: Effective visual feedback in gait rehabilitation. *IEEE Transactions on Human-Machine Systems* 48, 6 (2018), 626–636.
[49] Seven. 2020. Seven. https://seven.app/.
[50] Lior Shmuelof, John W Krakauer, and Pietro Mazzoni. 2012. How is a motor skill learned? Change and invariance at the levels of task success and trajectory control. *Journal of neurophysiology* 108, 2 (2012), 578–594.
[51] Roland Sigrist, Georg Rauter, Robert Riener, and Peter Wolf. 2013. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review. *Psychonomic bulletin & review* 20, 1 (2013), 21–53.
[52] Diego Silang Maranan, Sarah Fdili Alaoui, Thecla Schiphorst, Philippe Pasquier, Pattarawut Subyen, and Lyn Bartram. 2014. Designing for movement: evaluating computational models using LMA effort qualities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 991–1000.
[53] Rajinder Sodhi, Hrvoje Benko, and Andrew Wilson. 2012. LightGuide: projected visualizations for hand movement guidance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 179–188.
[54] Maurício Sousa, João Vieira, Daniel Medeiros, Artur Arsenio, and Joaquim Jorge. 2016. SleeveAR: Augmented reality for rehabilitation using realtime feedback. In *Proceedings of the 21st international conference on intelligent user interfaces*. ACM, New York, NY, USA, 175–185.

[55] Gabriele Spina, Guannan Huang, Anouk Vaes, Martijn Spruit, and Oliver Amft. 2013. COPDTrainer: a smartphone-based motion rehabilitation training system with real-time acoustic feedback. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, New York, NY, USA, 597–606.

[56] Thomas Stütz, Michael Domhardt, Gerlinde Emsenhuber, Daniela Huber, Martin Tiefengrabner, Nicholas Matis, and Simon Ginzinger. 2017. An interactive 3D health app with multimodal information representation for frozen shoulder. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, New York, NY, USA, 1–11.

[57] Richard Tang, Xing-Dong Yang, Scott Bateman, Joaquim Jorge, and Anthony Tang. 2015. Physio@ Home: Exploring visual guidance and feedback techniques for physiotherapy exercises. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 4123–4132.

[58] Atima Tharatipyakul, Kenny TW Choo, and Simon T Perrault. 2020. Pose Estimation for Facilitating Movement Learning from Online Videos. In *Proceedings of the International Conference on Advanced Visual Interfaces*. ACM, New York, NY, USA, 1–5.

[59] Yoshimasa Tokuyama, RPC Janaka Rajapakse, Sachiyo Yamabe, Kouichi Konno, and Yi-Ping Hung. 2019. A Kinect-Based Augmented Reality Game for Lower Limb Exercise. In *2019 International Conference on Cyberworlds (CW)*. IEEE, Long Beach, CA, USA, 399–402.

[60] Milka Trajkova and Francesco Cafaro. 2018. Takes Tutu to ballet: designing visual and verbal feedback for augmented mirrors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–30.

[61] Milka Trajkova and Mexhid Ferati. 2015. Usability evaluation of kinect-based system for ballet movements. In *International Conference of Design, User Experience, and Usability*. Springer, 464–472.

[62] Jeroen JG Van Merriënboer and John Sweller. 2010. Cognitive load theory in health professional education: design principles and strategies. *Medical education* 44, 1 (2010), 85–93.

[63] Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2013. Motionma: motion modelling and analysis by demonstration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1309–1318.

[64] Thomas Waltemate, Felix Hülsmann, Thies Pfeiffer, Stefan Kopp, and Mario Botsch. 2015. Realizing a low-latency virtual reality environment for motor learning. In *Proceedings of the 21st ACM symposium on virtual reality software and technology*. ACM, New York, NY, USA, 139–147.

[65] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, New York, NY, USA, 154–162.

[66] Graham Wilson and Stephen A Brewster. 2016. Using Dynamic Audio Feedback to Support Peripersonal Reaching in Young Visually Impaired People. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 209–218.

[67] Wenjie Wu and Stefan Rank. 2015. Responsive environmental Diegetic audio feedback for hand gestures in audio-only games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. ACM, New York, NY, USA, 739–744.

[68] Masahiro Yamaguchi, Shohei Mori, Peter Mohr, Markus Tatzgern, Ana Stanescu, Hideo Saito, and Denis Kalkofen. 2020. Video-Annotated Augmented Reality Assembly Tutorials. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 1010–1022.

[69] Ungyeon Yang and Gerard Jounghyun Kim. 2002. Implementation and evaluation of "just follow me": An immersive, VR-based, motion-training system. *Presence: Teleoperators & Virtual Environments* 11, 3 (2002), 304–323.

[70] Alexander D Young, Martin J Ling, and Damal K Arvind. 2011. IMUSim: A simulation environment for inertial sensing algorithm design and evaluation. In *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*. IEEE, Long Beach, CA, USA, 199–210.

[71] Shibo Zhang and Nabil Alshurafa. 2020. Deep generative cross-modal on-body accelerometer data synthesis from videos. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. ACM, New York, NY, USA, 223–227.

[72] Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 686–701.

[73] Yuhang Zhao, Elizabeth Kupferstein, Hathaitorn Rojnirun, Leah Findlater, and Shiri Azenkot. 2020. The effectiveness of visual and audio wayfinding guidance on smartglasses for people with low vision. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. ACM, New York, NY, USA, 1–14.

[74] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, CA, USA, 10394–10403.