SmartPTA: A Smartphone-Based Human Motion Evaluation System

Geonho Cha, Joonsig Gong, and Songhwai Oh

***Abstract*—This paper considers a human motion evaluation system (HMES) using two smartphones and a server. An HMES user wears retro-reflective markers on her 13 joints and follows a motion demonstrated by an expert. Two smartphones detect markers and transmit detected 2D marker positions to the server which reconstructs 3D motion of the user. The server performs a spatio-temporal alignment between the motion of the user and the motion demonstrated by an expert to evaluate the quality of the user's motion. The proposed system is applicable for physical therapy and sports education as an inexpensive alternative. We demonstrate that the 3D motion of the HMES user can be reconstructed reliably in real-time and successfully aligned with the reference motion of an expert, providing real-time feedback to the user.**

***Index Terms*—Motion evaluation systems, Motion capture, Smartphones, Real-time systems, 3D reconstruction**

I. Introduction

Following an action by watching the demonstration of an expert is the key for learning new activities [1]. But it is hard to get feedback about the learner's motion in real-time unless a teacher is present with the learner. A human motion evaluation system (HMES) can provide feedback to the learner or the user in the absence of a teacher or an expert. An HMES has a number of applications, such as physical therapy, sports education, and games.

A number of human motion evaluation systems have been proposed recently. Ghasemzadeh and Jafari [2] proposed a baseball swing evaluation system using a body sensor network. A user wears sensors on her chest, right wrist, and hip and each sensor consists of a three-axis accelerometer and a two-axis gyroscope. The motion of the user is analyzed and a feedback is given to the user. Spelmezan and Borchers [3] demonstrated a real-time snowboard training system. The proposed system measures the weight distribution on the snowboard using force-sensitive resistors and provides an immediate audio or tactile feedback to the user when an incorrect weight distribution is detected. Visual information can be useful for improving an HMES. Kwon and Gross [1] developed a motion training system for Taekwondo, a Korean martial art, which combines visual and body sensor data. Chan et al. [4] proposed a virtual-reality dance training system. A user wears a motion capture suit and imitates the motion demonstrated by a virtual teacher projected on the wall. The motion of the user is analyzed by the system and a feedback is given to the user. Chua et al. [5] proposed a virtual-reality motion training system for Tai Chi, a Chinese martial art. A user wears a motion capture suit and a wireless head mounted display (HMD) for viewing a rendered virtual environment. The user learns Tai Chi based on the feedback given by the system.

In computer vision, a number of methods have been developed to estimate a human pose from 2D images. These methods can be classified into two groups depending on whether markers are used or not. In [6], silhouettes were extracted from images and the likelihood of joint angles was computed. The human pose was constructed by estimating joint angles using the maximum a posteriori criterion. In [7], a human pose was tracked by fitting articulated surfaces to 3D cloud points and surface normals using the expectation maximization algorithm. In [8], voxels were computed from multiple images and skeleton poses were estimated from voxels. A skeleton model was fitted to voxels by aligning each bone of a skeleton to a corresponding set of voxels. However, a markerless motion capture method has high computational cost, which is not appropriate for a real-time system. In [9], four colored markers were used for extracting joints from two cameras. The locations of other joints were estimated using detected marker positions and a silhouette of the subject. In [10], a data-driven 3D human pose reconstruction method was presented. It detected a small number of retro-reflective markers using two cameras and reconstructed the full human pose using related poses in the database. In [11], a camera sensor network based 3D human pose reconstruction method was proposed.

In this paper, we are interested in a low-cost, easily-accessible human motion evaluation system based on a vision-based human pose estimation method. It consists of two smartphones and a notebook computer which are easy to access in daily life. Because of the low computational power of a smartphone, we adopted the marker based human pose

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2013R1A1A2065551).

G. Cha, J. Gong, and S. Oh are with CPSLAB and ASRI, Department of Electrical and Computer Engineering, Seoul National University, Seoul, Korea (e-mail: {geonho.cha, joonsig.gong, songhwai.oh}@cpslab.snu.ac.kr).

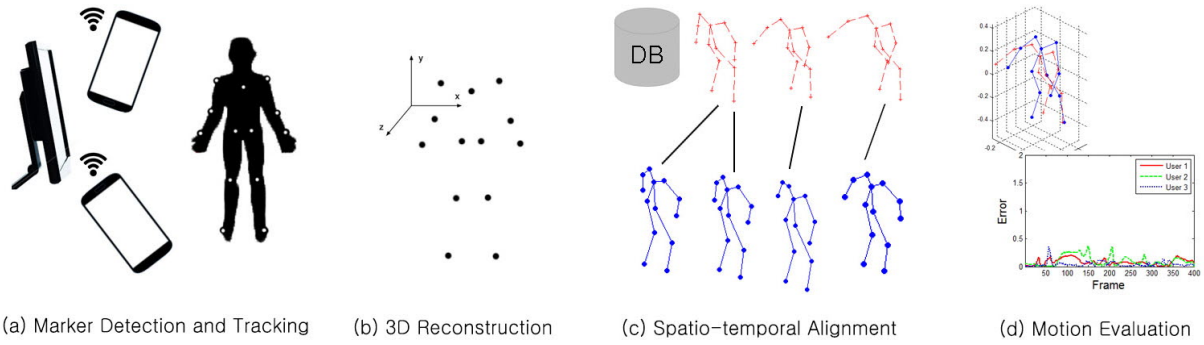


Fig. 1. An overview of the proposed system. It consists of four major modules: marker detection and tracking, 3D reconstruction, spatio-temporal alignment, and motion evaluation.

estimation method proposed in [11]. A user wears retro-reflective markers on her 13 joints and follows the motion of an expert in front of smartphones. The smartphones detect markers and report 2D joint positions to a server, a notebook computer, which reconstructs the 3D pose of the user. The server performs a spatio-temporal alignment between 3D motion sequences of the user and the expert to evaluate how well the user is performing.

The remainder of this paper is organized as follows. Section II describes the overview of the proposed human motion evaluation system using smartphones. Section III details the spatio-temporal alignment. Experimental results are provided in Section IV.

II. Human Motion Evaluation System using Smartphones

An overview of the proposed system is shown in Figure 1. There are four major modules in the system: (1) 2D joint marker detection, (2) 3D motion reconstruction, (3) a spatio-temporal alignment between the user's motion and the expert's motion, and (4) evaluation of the user's motion.

The proposed system consists of two smartphones and a notebook computer as a server (see Figure 2). While more smartphones can be used to improve the accuracy, in this paper, two smartphones are used since it is the minimum number of cameras required for 3D reconstruction of a scene. We have connected the server to a display to show reconstructed 3D motion in real-time. Smartphone cameras are calibrated using the method described in [12]. On a smartphone, a client application is running to detect 2D marker positions of 13 joints of the subject (see Figure 3). Each smartphone transmits detected 2D marker positions to the server using Wi-Fi at 15 frames per second.

1) Expert Motion Database: We assume that the system has a database of 3D motion sequences and videos of motions performed by experts. The expert motion database can be collected using the proposed system with the expert mode. An expert wears retro-reflective markers on her 13 joints and performs the motion in front of two calibrated cameras. The system reconstructs 3D motion sequences of the expert and stores them in the database. Examples of expert motions are shown in Figure 4. In each snapshot, a stick figure on the right is the reconstructed 3D pose of the expert. In our experiments, five motions are collected and they are two upper body motions (arm lifting and upper body stretching), two

Fig. 2. A photo of a user testing the proposed system. The proposed system consists of two smartphones and a notebook. In our experiments, a notebook is connected with a display to show the reconstructed 3D motion in real-time.

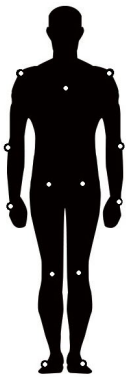
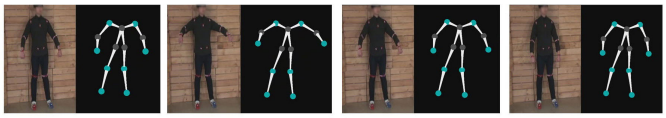
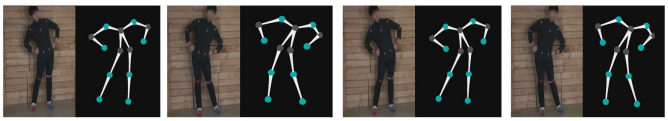


Fig. 3. A 13-joint human skeleton model.



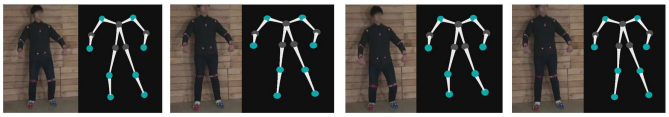
(a) Arm Lifting



(b) Upper Body Stretching

(c) Leg Stretching 1

(d) Leg Stretching 2



(e) Squat

Fig. 4. Examples from the expert motion database.

lower body motions (leg stretching 1 and leg stretching 2), and one whole body motion (squat) to demonstrate the feasibility of the proposed system. More expert motions can be easily added to the database.

2) Clients and Server: In our experiments, two Samsung Galaxy S2 smartphones (1.2GHz dual core CPU with 1GB RAM running Android 4.0.4) are used as clients. Using the camera of a smartphone, 2D positions of 13 joints of the user are detected from an image with resolution 800×450. In order to run the system in real-time, it is required to accelerate the marker detection process. We were able to process 15 frames per second for marker detection on a smartphone by running the simple marker detection algorithm from [11] on a GPU using the Android RenderScript library. The detected 2D marker positions are transmitted to the server via Wi-Fi.

A notebook computer is the server in our system, which receives 2D marker positions of joints from two clients. When the server receives two sets of detected 2D joint positions, it performs 3D human pose estimation based on the extrinsic camera parameters found from calibration using the method from [11]. We assume that two cameras of smartphones are calibrated prior to the use of the system based on the method described in [12]. Since the reconstructed 3D motion is aligned to the camera coordinate system, it is necessary to transform the coordinate system to the coordinate system used in the expert motion sequence before the spatio-temporal alignment.

In order to properly match the user's motion to the motion of the expert, a process of the spatio-temporal alignment is necessary since body dimensions of individuals are different and the same motion cannot be repeated exactly. Section III describes an algorithm for the spatio-temporal alignment. Once motions are aligned, the server evaluates the quality of the motion by the user using the distance between the aligned 3D motion of the user and the motion of the expert.

III. Spatio-Temporal Alignment

Since body dimensions of the expert and the user are different, it is necessary to align the 3D pose of the user to the pose of the expert considering differences in scale, rotation, and speed. The differences in scale and rotation can be solved using generalized Procrustes analysis [13], [14] and the difference in speed can be addressed using the dynamic time warping algorithm [15].

A. Generalized Procrustes Analysis

Generalized Procrustes analysis (GPA) is used to align shapes to a reference shape using rigid transformations [13]. Consider a set of m shapes, $\mathbf{S}_i \in \mathbb{R}^{d \times n}, i = 1, ..., m$, where each shape $\mathbf{S}_i$ contains positions of n joints $x_j \in \mathbb{R}^d, j = 1, ..., n$. GPA superimposes all m shapes to their mean shape $\bar{\mathbf{S}}$ by finding translations, rotations, and scale factors [13]. The problem can be formulated as the minimization of shape differences between $\mathbf{S}_i$ and $\bar{\mathbf{S}}$ with respect to rotations $\mathbf{R}_i$ and scale factors ρ_i, if we translate all shapes $\mathbf{S}_i$'s to have the origin $[0, ..., 0]^T$ as the common center.

$$\begin{aligned} &\arg\min_{\mathbf{R}_i, \rho_i} \sum_{i=1}^{m} \|\rho_i \mathbf{R}_i \mathbf{S}_i - \bar{\mathbf{S}}\|_F^2 \\ &\text{s.t.} \qquad \mathbf{R}_i^T \mathbf{R}_i = \mathbf{I}, \quad f(\rho) = 1, \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm for matrices, *i.e.*, $\|\mathbf{A}\|_F^2 = tr(\mathbf{A}^T\mathbf{A})$, and $\rho = [\rho_1, ..., \rho_m]^T$. $f(\rho) = 1$ is the constraint on the scale factors for preventing a trivial solution. For our problem, two motion shapes $\mathbf{S}_1, \mathbf{S}_2 \in \mathbb{R}^{3 \times n}$ are given, where $\mathbf{S}_1$ is the 3D motion shape of the expert and $\mathbf{S}_2$ is the 3D motion shape of the user. We use the scale factor constraint $f(\rho) = \text{vec}(\rho \mathbf{R} \mathbf{S}_2)^T \text{vec}(\mathbf{S}_1)$ inspired by [16], where $\text{vec}(\cdot)$ is a vectorization operator. If we normalize $\mathbf{S}_1$ and $\mathbf{S}_2$, such that $\|\mathbf{S}_1\|_F = 1$ and $\|\mathbf{S}_2\|_F = 1$, we can superimpose the 3D motion shape of the user to the expert's motion shape by calculating the rotation $\mathbf{R}$ and the scale factor ρ as follows [16]:

$$\mathbf{R} = \mathbf{V}^T \mathbf{U}, \tag{2}$$

where $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ is the singular value decomposition of $\mathbf{S}_2\mathbf{S}_1$, and

$$\rho = \frac{1}{tr(\mathbf{\Lambda})}. \tag{3}$$

Hence, we can align the 3D motion shape of the user $\mathbf{S}_2$ to the expert's motion shape $\mathbf{S}_1$ as follows:

$$\mathbf{S}_2^{'} = \rho\mathbf{R}\mathbf{S}_2. \tag{4}$$

B. Dynamic Time Warping

Dynamic time warping (DTW) is a popular algorithm to find the optimal alignment between two time-dependent sequences [15]. Given two time sequences $T_1 := \{\mathbf{S}_{11}, \mathbf{S}_{12}, ..., \mathbf{S}_{1N}\}$ and $T_2 := \{\mathbf{S}_{21}, \mathbf{S}_{22}, ..., \mathbf{S}_{2M}\}$, where $\mathbf{S}_{ij} \in \mathbb{R}^{3\times n}$ is a 3D motion shape, and a local cost measure $c(\mathbf{S}_1, \mathbf{S}_2) : \mathbb{R}^{3\times n} \times \mathbb{R}^{3\times n} \rightarrow \mathbb{R}_{\geq 0}$, the goal is to find the optimal alignment path p^* of length L, $p^* = (p_1, ..., p_L)$, where $p_i \in \mathbb{N}^2$, between T_1 and T_2 with the minimum overall cost. The alignment $p_i = (u, v)$ at time i assigns the u-th 3D shape in T_1 to the v-th 3D shape in T_2. To solve this problem, we define the accumulated cost matrix $\mathbf{D} \in \mathbb{R}^{N\times M}$ as follows:

$$\mathbf{D}_{(n,m)} = \begin{cases} \sum_{k=1}^{n} c(\mathbf{S}_{1k}, \mathbf{S}_{21}) \text{ if } n \in \{1, ..., N\}, m = 1, \\ \sum_{k=1}^{m} c(\mathbf{S}_{11}, \mathbf{S}_{2k}) \text{ if } n = 1, m \in \{1, ..., M\}, \\ \min\{\mathbf{D}_{(n-1,m-1)}, \mathbf{D}_{(n-1,m)}, \mathbf{D}_{(n,m-1)}\} \\ \quad + c(\mathbf{S}_{1n}, \mathbf{S}_{2m}) \qquad \text{else.} \end{cases} \tag{5}$$

$\mathbf{D}_{(n,m)}$ is the overall cost of an optimal path between $T_1' = \{\mathbf{S}_{11}, \mathbf{S}_{12}, ..., \mathbf{S}_{1n}\}$ and $T_2' = \{\mathbf{S}_{21}, \mathbf{S}_{22}, ..., \mathbf{S}_{2m}\}$, which are subsequences of T_1 and T_2, respectively. In our system, the following distance between two aligned 3D motion shapes is used as the cost measure:

$$c(\mathbf{S}_{1u}, \mathbf{S}_{2v}) = \|\mathbf{S}_{1u} - \rho\mathbf{R}\mathbf{S}_{2v}\|_F, \tag{6}$$

where ρ and $\mathbf{R}$ are computed using GPA as described in Section III-A. When a shape is matched to multiple shapes in other sequence, the minimum cost value is used.

Given the accumulated cost matrix $\mathbf{D}$, we can find the optimal path $p^* = (p_1, ..., p_L)$, which is computed in the reverse order starting with $p_L = (N, M)$ as follows [15]:

$$p_{t-1} := \begin{cases} (1, m-1) & \text{if } n = 1 \\ (n-1, 1) & \text{if } m = 1 \\ \arg\min(\mathbf{D}_{(n-1,m-1)}, & \\ \quad \mathbf{D}_{(n-1,m)}, \mathbf{D}_{(n,m-1)}) & \text{else.} \end{cases} \tag{7}$$

IV. Experiments

We have tested the proposed system with various motions and different users to evaluate the feasibility of the system. There is a total of five motions, consisting of arm lifting, upper body stretching, leg stretching 1, leg stretching 2, and squat as shown in Figure 4. Each motion is about 400 frames long, but the longer motions also could be used. The system is tested by four individuals: one expert and three learners.

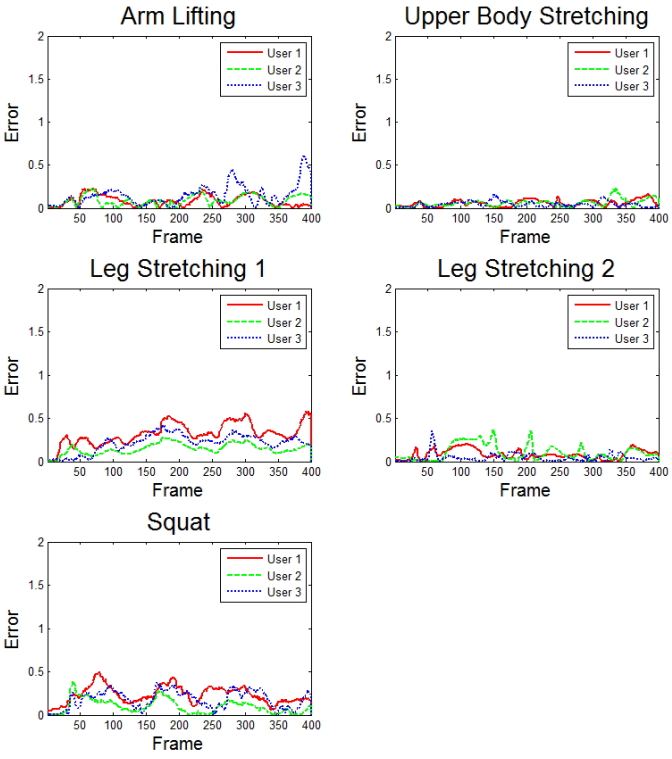


Fig. 5. Reconstruction errors of three users following five reference motions.

We showed the motion of the expert to a user using a display and asked the user to follow the motion of the expert. Three subjects followed the motion of the expert and Figure 5 shows reconstruction errors of users' motions with respect to the expert's motion. The reconstruction error is the relative distance between the aligned 3D motion shapes of the user and the expert computed based on the cost measure given in (6). In Figure 6, we show the reconstructed and aligned user poses along with the reference poses of the expert. Figure 7 shows some examples with large reconstruction errors. For the Arm Lifting motion, User 1 should have lifted the arm at a different angle, User 2 should have lifted the arm more, and User 3 lifted the left arm too much. For the Upper Body Stretching motion, all users should have put their hands on their waist. For Leg Stretching 1, users should have lifted their leg more. For Leg Stretching 2, users should have put their hands on waist. For Squat, the users should have lifted their arms more. The experiment verifies that the reconstruction error provides a meaningful information about how well a user follows the expert's motion.

Once users are familiar with the expert's motion, we have asked the users to follow the expert motion as closely as possible and computed the average reconstruction error between the user's motion and the expert's motion as shown in Figure

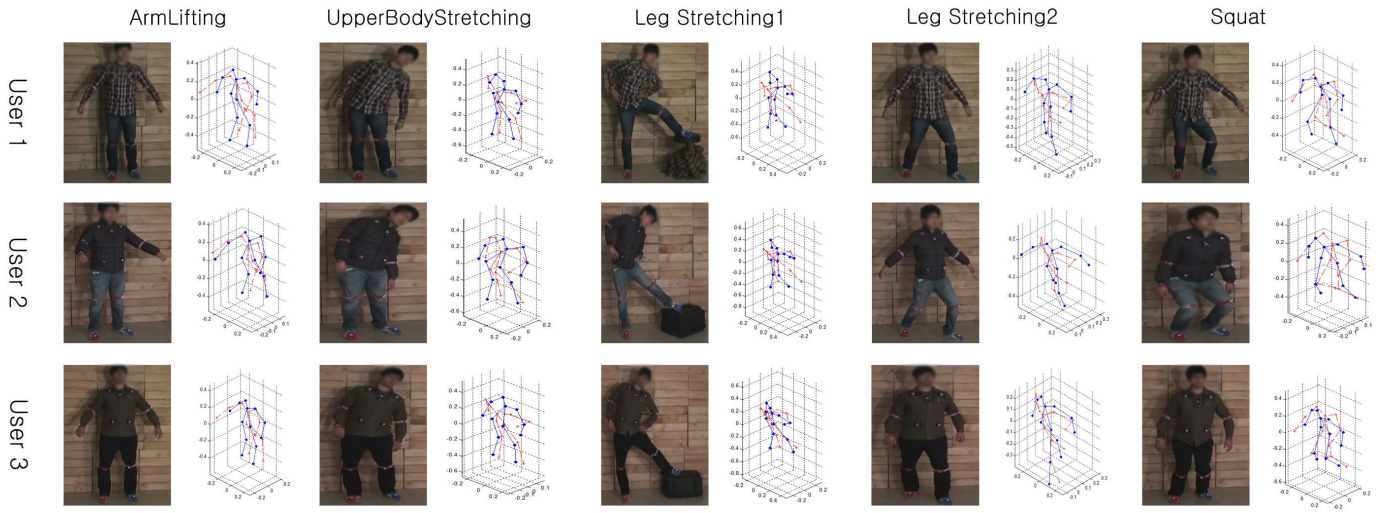


Fig. 6. Reconstructed 3D motions of users aligned to expert's motions. A red colored skeleton is the motion of an expert and a blue colored skeleton is the motion of a user.

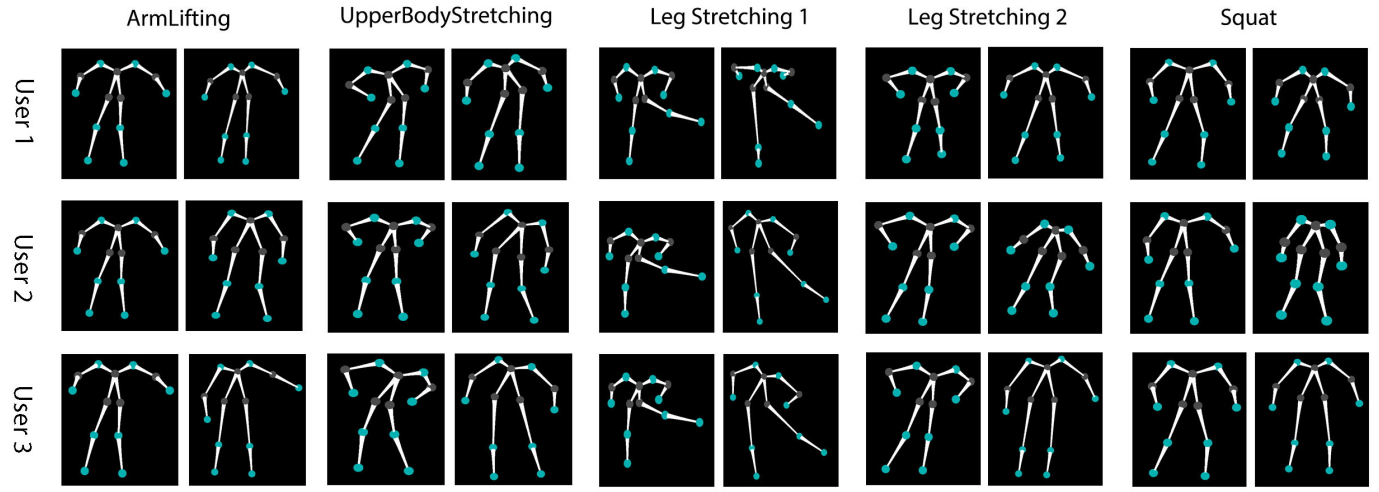


Fig. 7. Examples with large reconstruction errors. For each pair of stick figures, the left stick figure is the reference pose by an expert and the right stick figure is the estimated pose of a user.

8. The average error was about 0.55. Although we perform the spatio-temporal alignment between 3D motion sequences using GPA and DTW, there can exist a bias between two 3D motion sequences since the exact alignment is not always possible and there will be a nonzero reconstruction error.

A. Feedback to the User

The proposed human motion evaluation system can provide a feedback to the user showing how well he or she is following the motion of the expert with a percentage score right after the user's practice. We converted the reconstruction error into a percentage considering the average reconstruction error shown in Figure 8. Examples of this user feedback demonstration are shown in Figure 9. Once a user completes a motion sequence, the user can replay his or her 3D reconstructed motion along with the expert's motion to obtain a further feedback.

V. Conclusions

We have shown that smartphones can be used to develop a human motion evaluation system. We have demonstrated that the 2D markers of 13 joints can be reliably tracked in real-time using smartphones. In order to evaluate the motion of the user, we have reconstructed the 3D motion of the user and aligned the reconstructed motion to the reference expert's motion using GPA and DTW. We were able to successfully provide a score to the user in real-time, such that the user can follow the expert's motion better. We envision that a number of useful smartphone applications can be developed based on the proposed framework.

References

[1] D. Y. Kwon and M. Gross, "Combining body sensors and visual sensors for motion training," in *Proc. of the ACM SIGCHI International*

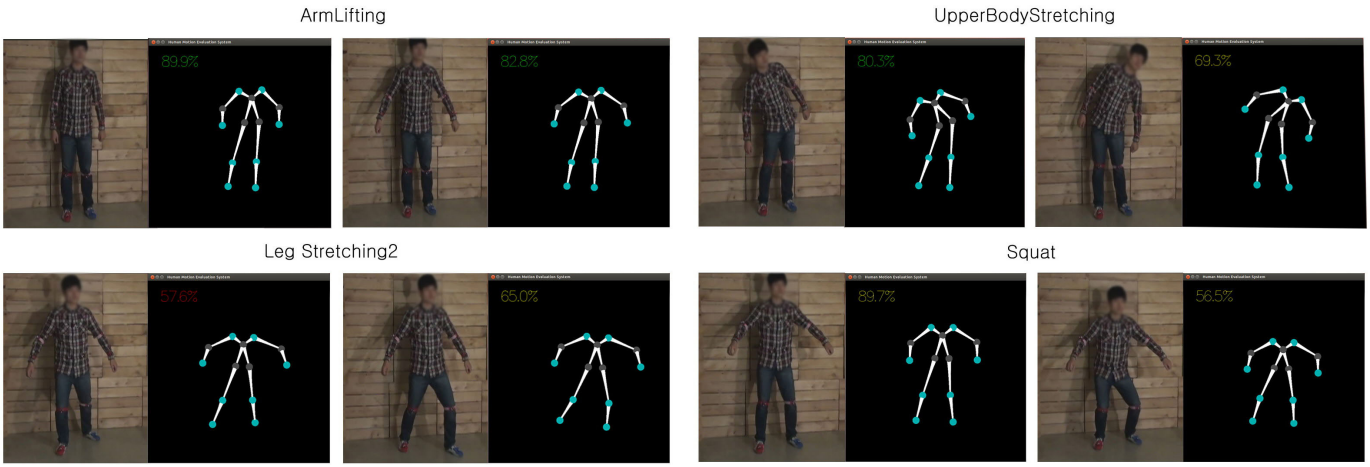


Fig. 9. Screenshots of the system providing a feedback to the user.

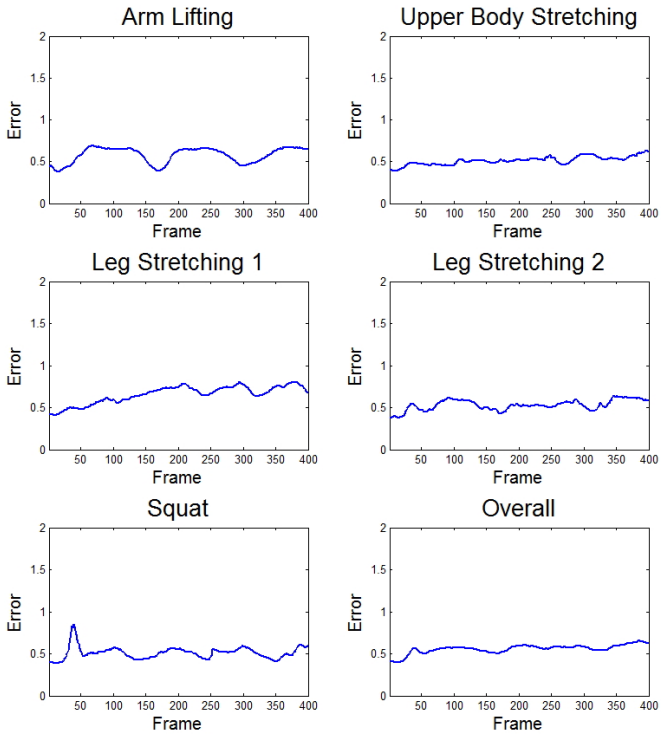


Fig. 8. Average reconstruction errors of five motions when the users follow each motion as closely as possible. The figure titled Overall shows the average error of five motions.

Conference on Advances in Computer Entertainment Technology, 2005.

[2] H. Ghasemzadeh and R. Jafari, "Coordination analysis of human movements with body sensor networks: A signal processing model to evaluate baseball swings," *Sensors*, vol. 11, no. 3, pp. 603–610, 2011.

[3] D. Spelmezan and J. Borchers, "Real-time snowboard training system," in *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, 2008.

[4] J. C. Chan, H. Leung, J. K. Tang, and T. Komura, "A virtual reality dance training system using motion capture technology," *IEEE Transactions on Learning Technologies*, vol. 4, no. 2, pp. 187–195, 2011.

[5] P. T. Chua, R. Crivella, B. Daly, N. Hu, R. Schaaf, D. Ventura, T. Camill, J. Hodgins, and R. Pausch, "Training for physical tasks in virtual environments: Tai Chi," in *Proc. of the IEEE Conference on Virtual Reality*, 2003.

[6] C. Sminchisescu and A. Telea, "Human pose estimation from silhouettes: A consistent approach using distance level sets," in *Proc. of the International Conference on Computer Graphics, Visualization and Computer Vision*, 2002.

[7] R. Horaud, M. Niskanen, G. Dewaele, and E. Boyer, "Human motion tracking by registering an articulated surface to 3d points and normals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 158–163, 2009.

[8] A. Sundaresan and R. Chellappa, "Markerless motion capture using multiple cameras," in *Proc. of the IEEE Conference on the Computer Vision for Interactive and Intelligent Environment*, 2005.

[9] H. Ukida, S. Kaji, Y. Tanimoto, and H. Yamamoto, "Human motion capture system using color markers and silhouette," in *Proc of the IEEE Conference on Instrumentation and Measurement Technology Conference*, 2006.

[10] J. Chai and J. K. Hodgins, "Performance animation from low-dimensional control signals," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 686–696, 2005.

[11] H. Oh, G. Cha, and S. Oh, "Samba: A real-time motion capture system using wireless camera sensor networks," *Sensors*, vol. 14, no. 3, pp. 5516–5535, 2014.

[12] S. Yoon, H. Oh, D. Lee, and S. Oh, "PASU: A personal area situation understanding system using wireless camera sensor networks," *Personal and Ubiquitous Computing*, vol. 17, no. 4, pp. 713–727, 2013.

[13] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[14] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 285–339, 1991.

[15] M. Müller, "Dynamic time warping," *Information Retrieval for Music and Motion*, pp. 69–84, 2007.

[16] M. Lee, J. Cho, C.-H. Choi, and S. Oh, "Procrustean normal distribution for non-rigid structure from motion," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.