3D Tracking Using Smartphones for a Marker-Based Optical Motion Capture System

Kengo Teramoto[1], Ryo Mukainakano[1], Hiroki Watanabe[1], Hiromichi Hashizume[2], and Masanori Sugimoto[1]

[1] Hokkaido University, Sapporo 060-0814, Japan
{kteramoto, mukainakano, hiroki.watanabe, sugi}@ist.hokudai.ac.jp
[2] National Institute of Informatics, Tokyo, 101-8430, Japan
has@nii.ac.jp

Abstract. Our research group is developing a marker-based optical motion capture system using smartphones in a mobile setting. The proposed system can be realized with low cost, high accuracy of measurement, and can be easily deployed. This paper describes a 3D tracking technique using smartphones, which is one of the key parts of the proposed motion capture system. The system requires multiple smartphones, an LED light source, and speakers. The 3D coordinates of the cameras are obtained by acoustic localization, and the 3D coordinates of the markers are obtained by the principle of triangulation from the obtained images and the 3D coordinates of the smartphones. 3D tracking experiments using two smartphones in fixed settings confirmed that the proposed system could achieve millimeter-level accuracy. Its 3D tracking performance in a handheld setting was also tested.

Keywords: 3D tracking, Motion capture, Smartphone

1 Introduction

A motion capture system (MCS) is used in scenarios such as sports science, medical research, the creation of CG models in games, and movies. Among them, a marker-based optical MCS is the most popular MCS because it can measure with high accuracy. However, the commercialized optical MCS is extremely expensive at a cost ranging from tens to hundreds of thousands of US dollars. In addition, it is necessary to accurately measure the positional relationship between multiple cameras, requiring high deployment costs. As a result, such systems are usually used in a fixed nonmobile setting. In contrast, smartphones have become ubiquitous in today's world. Smartphones contain different types of sensors such as cameras, speakers, microphones, and inertial sensors, and they can be used for various applications. Our research group is developing an MCS using smartphones, thereby making the system mobile. The proposed system can be realized at a low cost because it employs users' smartphones, an LED light source, and three loudspeakers. Moreover, the proposed system can measure the positions of markers with high accuracy and can be easily deployed without manually measuring the positional relationship between cameras. This paper evaluated the

performance of one of the key technologies of the proposed system, 3D tracking, and confirmed that it could achieve millimeter-level accuracy in estimating the 3D positions of moving markers. The system was also tested in a handheld setting to verify its 3D tracking performance.

2 Related Work

There are various types of MCSs; those using cameras are called optical MCSs and comprise two major types. One is marker-based (e.g., [1]), in which the 3D coordinates of the markers attached to a human body are calculated by the principle of triangulation using images captured with multiple cameras. Some MCSs implementing high resolution and high frame rate cameras can accurately capture high-speed motions, and it is also possible to capture multiperson motions simultaneously in a wide indoor space. However, if the markers enter a blind spot, they cannot be detected. To alleviate this problem, the number of cameras must be increased. The other MCS type is markerless and does not require a marker to be attached to subjects (e.g., [2]). The regions of subjects are extracted from photographed images by using image recognition techniques, and their 3D movement is estimated by tracking them in real time. Although the markerless type has the advantage that it can reduce the physical burdens of subjects, its measurement accuracy does not achieve that of the marker-based MCS. The markerless MCS using smartphones with built-in camera was proposed by Wang et al. [3]. Through experiments on tracking a moving person, the average error in 500 frames was reported to be 4.23 cm; however, the authors did not discuss deployment issues related to smartphone localization. Another example of an MCS using smartphones was proposed by Pascu et al. [4]. In their system, smartphones are attached to various parts of a person's body to capture motions through the smartphone's built-in acceleration sensors.

3 Proposed System

3.1 System Overview

A marker-based optical type MCS was investigated to achieve higher accuracy in a mobile setting. The proposed system mainly performs three processes. The first process is to perform localization of smartphones and obtain their 3D positional relationship. The second process is to conduct the frame interpolation between smartphones so that captured images by different smartphone cameras are time-synchronized. The third process determines the 3D positions of markers to be attached to a person's body and captured by each smartphone using triangulation. This paper describes the current implementation status of the system using two smartphones.

3.2 Smartphone Localization

For the localization of the smartphones, the time-of-arrival (ToA) trilateration method [5] using optical and acoustic signals was used. To perform time synchronization between smartphones and a transmitter, optical and acoustic signals are

simultaneously transmitted from the transmitter. In addition, the signals are respectively received by a camera and a microphone built into the smartphone. The transmitter is composed of one LED light source and three loudspeakers. The frequency of the optical signal is set to one-third of the camera frame rate. The phase of the optical signal is detected by the three-point demodulation method [5] and is used as a reference point for time synchronization between the transmitter and smartphones. An acoustic signal emitted from each speaker is called a sync pattern having two sinusoidal waves with different frequencies and lasting 4 ms. A method called FDM-PAM [6] can accurately detect the signal reception time and estimate the distance between the speaker and smartphones. By conducting ToA trilateration using three speakers, the 3D positions of smartphones are identified.

3.3 Frame Interpolation

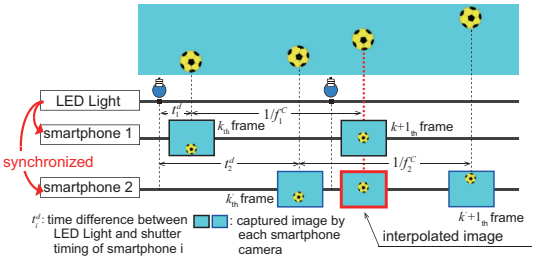

Fig. 1. Frame interpolation

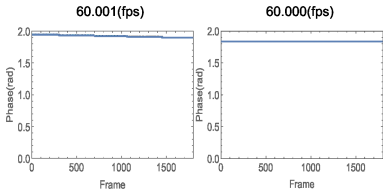

Fig. 2. Phase changes of a smartphone for 30 s (left: "without calibration", right: "with calibration")

Although a reference point of time synchronization for ToA trilateration is accurately identified by individual smartphone cameras as discussed in Section 3.2, their shutter timings are different from each other. In Fig. 1, differences between a signal emission time of an LED light and shutter timings of smartphones 1 and 2 are obtained as t_1^d and t_2^d, respectively. Therefore, a moving target is captured at different times by different smartphones, which produces localization errors. To solve the problem, a frame interpolation method is applied to obtain the target image at the same timing by different smartphones,

As shown in Fig. 1, a target is captured at the k-th frame by the smartphone 1 and $k^{'}$-th frame by the smartphone 2, and the difference between their shutter timings is $t_2^d - t_1^d$. Suppose that the positions of the target are given as $(u_{1,k}, v_{1,k})$ and $(u_{2,k'}, v_{2,k'})$ in the images captured by the smartphones 1 and 2, respectively. By conducting linear interpolation using the $k^{'}$-th and $(k^{'}+1)$-th frames of the smartphone 2, an interpolated target position $(u_{1,k+1}^{i2}, v_{1,k+1}^{i2})$ corresponding the $(k+1)$-th frame of the smartphone 1 is estimated as

$$u_{1,k+1}^{i2} = (u_{2,k'+1} - u_{2,k'}) f_2^C (\frac{1}{f_1^C} - (t_2^d - t_1^d))$$

$$v_{1,k+1}^{i2} = (v_{2,k'+1} - v_{2,k'}) f_2^C (\frac{1}{f_1^C} - (t_2^d - t_1^d)),$$

where f_1^C and f_2^C are camera frame rates of the smartphones 1 and 2, respectively.

3.4 Marker Localization

When internal and external parameters of cameras are known, the 3D coordinates of markers on the images captured by cameras are determined. The internal parameters are those related to focal length, image center, and distortion, and the external parameters are those related to rotation and translation. The internal parameter matrices A_1 and A_2 of two smartphone cameras are assumed to be known using the camera calibration method [7]. The external parameter matrices $[R_1|\boldsymbol{t_1}]$ and $[R_2|\boldsymbol{t_2}]$ of the smartphone cameras are composed of translational and rotational components, respectively. The former is related to the positions of the cameras and obtained through the localization method described in the previous section. The latter is related to the orientation and pose of the cameras and obtained by using built-in inertial sensors such as an accelerometer, gyroscope and magnetic sensor. It is also possible to use a visual marker attached at a neighboring area to the LED light (Fig. 3) and captured by the camera to increase the level of the estimation accuracy of the rotational components. By using the internal and external parameter matrices, perspective projection matrices are given as P_1 and P_2. Suppose (u_1, v_1) and (u_2, v_2) in the image plane of the two smartphones are a target point $\boldsymbol{Q_w} = (X_w, Y_w, Z_w)^T$ in the world coordinate. Then, the following equations (1) and (2) related to the 3D coordinates of the target point hold.

$$(p_{1,31}u_1 - p_{1,11})X_w + (p_{1,32}u_1 - p_{1,12})Y_w + (p_{1,33}u_1 - p_{1,13})Z_w = p_{1,14} - p_{1,34}u_1 \quad (1)$$

$$(p_{1,31}v_1 - p_{1,21})X_w + (p_{1,32}v_1 - p_{1,22})Y_w + (p_{1,33}v_1 - p_{1,23})Z_w = p_{1,24} - p_{1,34}v_1, \quad (2)$$

where P_i (i=1,2) is given as

$$P_i = \begin{bmatrix} p_{i,11} & p_{i,12} & p_{i,13} & p_{i,14} \\ p_{i,21} & p_{i,22} & p_{i,23} & p_{i,24} \\ p_{i,31} & p_{i,32} & p_{i,33} & p_{i,34} \end{bmatrix}. \quad (3)$$

$\boldsymbol{Q_w}$ is found by solving $\boldsymbol{b} = B\boldsymbol{Q_w}$, where B and $\boldsymbol{b}$ are given as Eqs. (4). The least squares solution of the target point $\hat{\boldsymbol{Q}}_{\boldsymbol{w}}$ is given by Eq. (5). B^+ is the pseudo-inverse of B, and this method can be applied when the number of cameras is three or more.

$$B = \begin{bmatrix} p_{1,31}u_1 - p_{1,11} & p_{1,32}u_1 - p_{1,12} & p_{1,33}u_1 - p_{1,13} \\ p_{1,31}v_1 - p_{1,21} & p_{1,32}v_1 - p_{1,22} & p_{1,33}v_1 - p_{1,23} \\ p_{2,31}u_2 - p_{2,11} & p_{2,32}u_2 - p_{2,12} & p_{2,33}u_2 - p_{2,13} \\ p_{2,31}v_2 - p_{2,21} & p_{2,32}v_2 - p_{2,22} & p_{2,33}v_2 - p_{2,23} \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} p_{1,14} - p_{1,34}u_1 \\ p_{1,24} - p_{1,34}v_1 \\ p_{2,14} - p_{2,34}u_2 \\ p_{2,24} - p_{2,34}v_2 \end{bmatrix} \quad (4)$$

$$\hat{\boldsymbol{Q}}_{\boldsymbol{w}} = B^+\boldsymbol{b}, B^+ = (B^TB)^{-1}B^T \quad (5)$$

4 3D Tracking Experiment

4.1 Clock Difference between Transmitter and Smartphones

Ideally, once the transmitter and smartphones are synchronized, the proposed method can conduct ToA trilateration for tracking smartphone positions by

using only acoustic signals as discussed in Section 3.2. However, this is not possible in practice. Fig. 2 shows the phase values of optical signals acquired by the left smartphone in 30-s measurements; it was confirmed that the phase value changed during the measurements. The working frame rate of the smartphone camera of 60.001 fps, despite having set the rate to exactly 60.0 fps, affected the phase value. As a result, the standard deviations of the smartphone localization in the measurements deteriorated, as shown in Table 1 ("without calibration").

Fig. 2 shows that the phase value changes at a constant rate. Thus, it is possible to conduct calibration by linear regression. In Table 1, "with calibration" means using a smartphone with the calibration by linear regression, and "FG synchronized" means that the optical signal generated by a function generator was accurately set to one-third of the smartphone camera's frame rate (60.001 fps). The table shows that the localization results of the smartphone calibrated by linear regression were improved to almost the same level as the localization results of the FG synchronized smartphone and that the calibration by linear regression proved to be effective. In our investigation, because actual frame rates set to 60 fps differed between smartphones, their frame rates should be measured and the linear regression for each smartphone should be conducted.

	X	Y	Z
with calibration	2.79	3.511	2.27
without calibration	14.63	26.19	35.73
FG synchronized	2.77	2.87	3.46

Table 1. Standard deviation of smartphone localization with and withoout calibration (mm)

Fig. 3. Experimental environment

4.2 Experimental Setup

The experimental setting is shown in Fig. 3. Two smartphones (iPhone 6s Plus) were used for tracking two color markers attached to a moving square bar under a fluorescent lamp in a room. A transmitter was placed at a line-of-sight position from the smartphones so they could capture transmitted optical and acoustic signals. Then, three speakers (FT200D, Fostex) mounted on the transmitter formed an equilateral triangle with a side length (baseline) of 300 mm, and the light source consisting of 56 white LEDs (OSW54L5111P, OptoSupply) was placed at the center of the triangle. The smartphone cameras had 1920 x 1080 pixels and were operated at 60 fps. Then, 20 Hz sinusoidal signals whose frequency was one-third of the camera frame rate were generated by a function generator (WF1948, NF Corporation) and emitted from the LED light. Sync patterns transmitted from the three speakers were composed of a pair of sinusoidal waves whose frequencies were 12.75 and 13.25 kHz, 13.75 and 14.25 kHz, and 14.75 and 15.25 kHz, respectively. As they were emitted every 0.2 seconds so the position update rate of the smartphones was 5 Hz. The linear regression

was conducted to compensate the clock difference between the smartphone as discussed in Section 4.1.

By receiving signals from the transmitter, the 3D position of each smartphone was calculated. Then, the position, pose, and captured image data were transferred from the smartphone to a PC via WiFi. The application software running on the smartphones was implemented using Swift 4. The 3D position estimation of markers was conducted using OpenCV 4.0 on the PC, and the measurements were carried out 10 times.

Two smartphone settings were tested in the experiments:

Experiment1 fixed on a tripod as shown in Fig. 3, and
Experiment2 held by a user.

The baseline between the two tripods was set to 0.39 m in Experiment 1 and users holding a smartphone stood at the same positions of the tripods in Experiment 2. The distance between the markers and smartphones ranged from 0.94 m to 1.04 m. Tracking results obtained by the proposed system were compared with those by a high-performance and commercially off-the-shelf motion capture system (MAC3D System, Motion Analysis Corporation) composed of 6 cameras (600 fps, 2048 x 1088 pixels) as the reference system. 3D tracking errors of the proposed system were given as differences of estimated marker 3D positions from the reference system.

4.3 Experiment 1: Fixed on a Tripod

Fig. 4, Fig. 5 and Table 2 show the tracking results. In Fig. 4 (a) and (b), the bar was translated quickly (captured frames: 27) and slowly (captured frames: 82) in the horizontal direction. As shown in Table 2, root means square errors (RMSEs) of the marker tracking were 7.15 mm (marker1, quick), 5.44 mm (marker2, quick), 2.20 mm (marker1, quick) and 1.93 mm (marker2, slow), respectively, without applying the frame interpolation method. The improvement of the tracking performance ranged from 0.01 to 1.99 mm with this method. Fig. 5 (a) and (b) show the results when the bar was translated quickly (captured frames: 21) and slowly (captured frames: 75) in the vertical direction. As shown in Table 2, RMSEs were 1.89 mm (marker1, quick), 2.33 mm (marker2, quick), 2.77 mm (marker1, quick) and 1.34 mm (marker2, slow), respectively, without applying the frame interpolation method. The improvement of the tracking performance ranged from 0.10 to 0.34 mm with the frame interpolation. Fig. 6 shows trajectories of the markers translated quickly and vertically (Fig. 5(a)). From the figure, the estimated trajectories are more unstable in the z-axis (depth) direction than in the x and y-axis directions. According to the theory of dilution of precision (DOP) [8] discussed in global navigation satellite system (GNSS) communities, the greater scattering along the z-axis was related to the shorter distance (baseline, 0.39 m) between the smartphones than that between the markers and the smartphones (0.97 - 1.04 m). By extending the baseline and increasing the number of smartphones, the localization results would be more stable and robust to the position estimation errors of smartphones, which tells

us a design guideline of the proposed system to fulfill a required level of the performance.

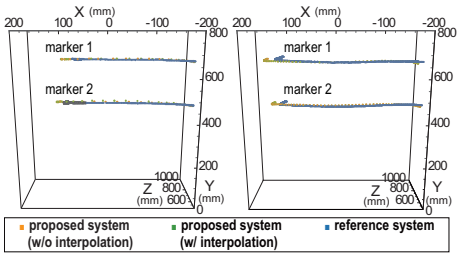

Fig. 4. Marker tracking when the bar was moved horizontally. (a) quick move (left), (b) slow move (right)

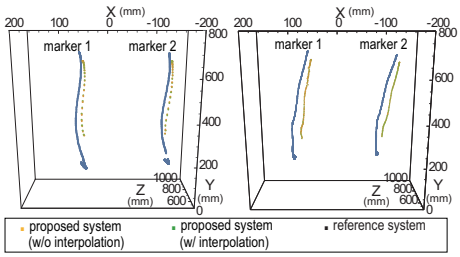

Fig. 5. Marker tracking when the bar was moved vertically. (a) quick move (left), (b) slow move (right)

		horizontal translation		vertical translation	
		w/o interpolation	w/ interpolation	w/o interpolation	w/ interpolation
quick	marker1	7.15	5.40	1.89	1.65
	marker2	5.44	3.45	2.33	2.01
slow	marker1	2.20	2.09	2.77	2.51
	marker2	1.93	1.87	1.34	1.24

Table 2. Average 3D tracking errors [mm]

4.4 Experiment 2: Held by a User

Each user was asked to stand still, and capture a single marker moving horizontally (captured frames: 51) using a handheld smartphone. The RMSE of the tracking was 168.28 mm without the frame interpolation, which was slightly improved to 153.79 mm with it. One of the causes of the performance deterioration was related to inaccurate 3D localization of the smartphones which were unstably held. The inaccuracy seemed significantly affected the marker tracking results. Further investigations are being conducted.

5 Conclusion and Future Work

This paper proposed a 3D tracking system using smartphones for a marker-based optical MCS in a mobile setting. The proposed system did not request the positions of the cameras to be set manually and could estimate them automatically by ToA trilateration using optical and acoustic signals. The experimental results showed that the proposed system could achieve 3D position estimation of markers with millimeter-level accuracy in fixed settings. A number of issues remain to be investigated in our future work. The next step is to extend the current system with more than two smartphones toward more accurate and precise motion-capture systems.

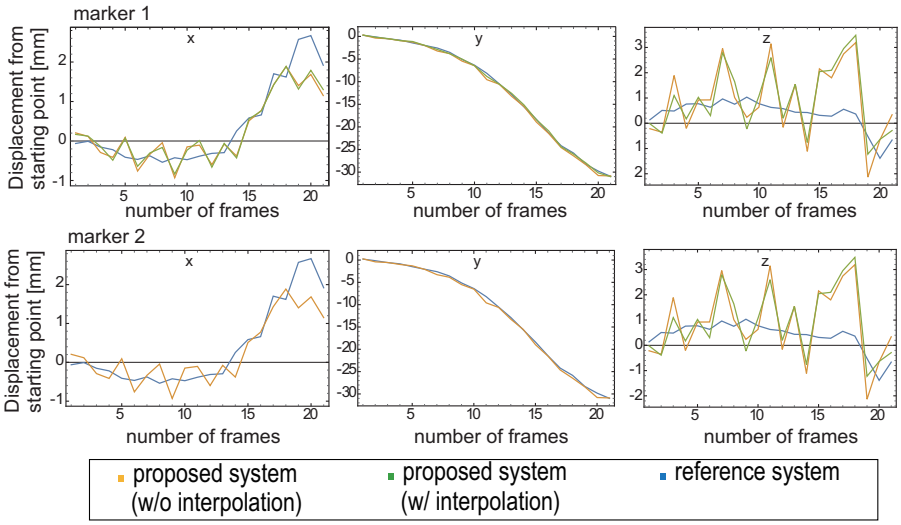

Fig. 6. Marker trajectories in the x, y, and z-axis directions

References

1. OptiTrack. "OptiTrack - Motion Capture Systems". OptiTrack. https://optitrack.com/, (accessed 2018-11-01).
2. Tracklab. "Organic Motion OpenStage 2.0 - Tracklab". Tracklab. https://tracklab.com.au/products/hardware/organic-motion-openstage-2-0/, (accessed 2018-11-01).
3. Y. Wang, Y. Liu, X. Tong, Q. Dai and P. Tan. "Outdoor Markerless Motion Capture with Sparse Handheld Video Cameras". *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 5, pp. 1856-1866, 2017.
4. T. Pascu, M. White, and Z. Patoli. "Motion Capture and Activity Tracking using Smartphone-driven Body Sensor Networks". In *Proceedings of INTECH 2013*, London, UK, pp.456-462, 2013.
5. T. Akiyama, M. Sugimoto, and H. Hashizume. "Time-of-arrival-based Smartphone Localization Using Visible Light Communication". In *Proceedings of IPIN 2017*, Sapporo, Japan, DOI:10.1109/IPIN.2017.8115904 , 2017.
6. M. Nakamura, T. Akiyama, M. Sugimoto, and H. Hashizume, "FDM-PAM: Rapid and Precise Indoor 3D Localization using Acoustic Signal for Smartphone". In *Proceedings of UbiComp 2014 Adjunct*, Seattle, WA, pp. 123-126, 2014.
7. Z. Zhang, "A flexible new technique for camera calibration". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 22, no. 11, pp. 1330-1334, 2000.
8. R.B. Langley, "Dilution of precision". *GPS World*, vol.10, no.5, pp.52-59, 1999.