A Unified Framework for Remote Collaboration Using Interactive AR Authoring and Hands Tracking

Jeongmin Yu, Jin-u Jeon, Gabyong Park, Hyung-il Kim, and Woontack Woo(✉)

KAIST UVR Lab, Daejeon, South Korea
{jmyu119,zkrkwlek,gypark,hyungil,wwoo}@kaist.ac.kr

Abstract. In this paper, we present a unified framework for remote collaboration using interactive augmented reality (AR) authoring and hand tracking methods. The proposed framework enables a local user to organize AR digital contents for making a shared working environment and collaborate multiple users in the distance. To develop the framework, we combine two core technologies: (i) interactive AR authoring method utilizing a smart input device for making a shared working space, (ii) hand-augmented object interaction method by tracking two hands in egocentric camera view. We implement a prototype of the proposed remote collaboration framework for testing its feasibility in an indoor environment. To the end, we expect that our framework enables collaboration as feeling a sense of co-presence with remote users in a user's friendly AR working space.

Keywords: Interactive AR authoring · Hand-augmented object interaction · Remote collaboration system

1 Introduction

Augmented reality (AR) is technology that enables users to close in supplementary information by seamlessly mixed with virtual objects in the real world [1]. Using this, the users can be worked with digital virtual elements and guided some needed directions. These useful information can be displayed in various devices such as mobile phones, PDA, head mounted display (HMD), and high performance PCs. AR technology is applied to various fields such as interactive games, education, military, gallery/exhibition, and repair/maintenance [2].

For the past few decades, AR applications have been mainly developed for only one user in the manner of one-way interaction with 3D virtual objects [5–7]. Even though they gives useful and interesting experience to the user, they do not provide experience of interaction and collaboration with other users. Recently, HMD-based remote collaboration systems have been developed to collaborate a shared target work with remote users [3, 4]. Unlikely existing remote collaboration systems [9, 10], these systems enable spatially un-limited interactions and give a sense of co-presence to the local user. However, these systems not only provides the confined simple interactions (e.g., flipping, grasping) by tracking a bare hand, but also provides a manually user-defined working environment to users.

N. Streitz and P. Markopoulos (Eds.): DAPI 2016, LNCS 9749, pp. 132–141, 2016.
DOI: 10.1007/978-3-319-39862-4_13

Meanwhile, many researchers have been studied on AR authoring systems for easily handling AR digital contents to users. For instance, [11, 12] have been attempted to AR authoring on mobile device. [11] shows interaction with AR contents using multi-touch interface of smart device. [12] presents an AR authoring method for unknown outdoor scene using mobile devices. However, because they do not generate a 3D map using depth sensors, they are unsuitable to register virtual digital contents on indoor environment. On the other hand, Project Tango [13] is a mobile authoring device that builds a 3D map of unknown indoor scene using a depth sensor. However, this system has some cumbersome points that a user spreads own arms enduringly during performing and sees the augmented spot through a narrow mobile device display.

In this paper, to settle above mentioned shortcomings, we present a novel HMD-based remote collaboration framework using interactive AR authoring and hand-augmented object interaction technologies. To develop the proposed framework, we integrate two main technologies which are interactive AR authoring with a wearable smart device (e.g., smartphone) for making a shared working space, and hand-augmented object interaction by tracking two bare hands in egocentric camera view. Through the proposed system, the local user can author his/her own working space easily without any professional programming skills [8], and collaborate remote users through intuitive interactions between tracking two hands and augmented objects. Through a preliminary prototype system implementation, we confirm its feasibility as a future remote collaboration platform. We expect that the proposed system can be applicable to many AR collaborative applications such as medical surgery education, urban planning, games and so forth.

The remainder of this paper is organized as follows. The proposed framework is presented in Sect. 2. In Sect. 3 introduces the initial implementation and preliminary result of the proposed framework. Lastly, the conclusions and outline plans for future works are presented in Sect. 4.

2 Proposed Framework

2.1 Overall Framework

Figure 1 shows the proposed overall system diagram of HMD-based remote collaboration. For this system, we use a smartphone for AR digital contents authoring, and use an egocentric short-range RGB-D camera and a wearable sensor (e.g., smartwatch) for accurate two hands tracking, and use an exocentric RGB-D camera for full-body tracking. For AR authoring, we use the positions, rotations and touch directions information from smartphone. For hands-augmented object interaction, we first segment bare hands from the egocentric camera. Then, hands and fingers are tracking based on a model fitting method. After registration between virtual and real hands, we can interact with a 3D augmented object for performing a shared target task. The detail methodological descriptions of interactive AR authoring and hands-augmented object interaction are presented as follows.

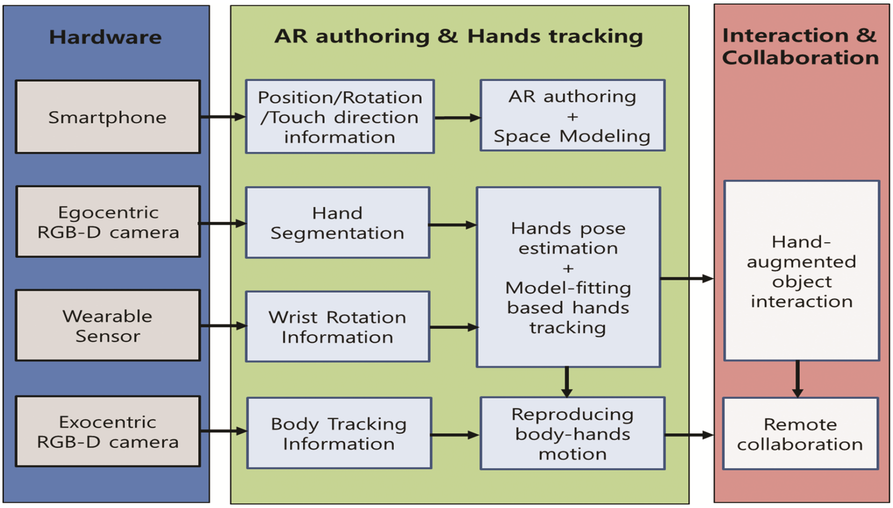
Hardware
Smartphone
Egocentric RGB-D camera
Wearable Sensor
Exocentric RGB-D camera
AR authoring & Hands tracking
Position/Rotation /Touch direction information
AR authoring + Space Modeling
Hand Segmentation
Hands pose estimation + Model-fitting based hands tracking
Wrist Rotation Information
Body Tracking Information
Reproducing body-hands motion
Interaction & Collaboration
Hand-augmented object interaction
Remote collaboration

Fig. 1. The proposed framework for HMD-based remote collaboration

2.2 Interactive AR Authoring

Figure 2 shows the pipeline of proposed interactive AR authoring system. We first compute the initial local reference coordinates of a target working space, and then these local reference coordinates transformed by the obtained simultaneous localization and mapping (SLAM)-based coordinates. AR digital contents/objects are placed in the transformed local coordinates.

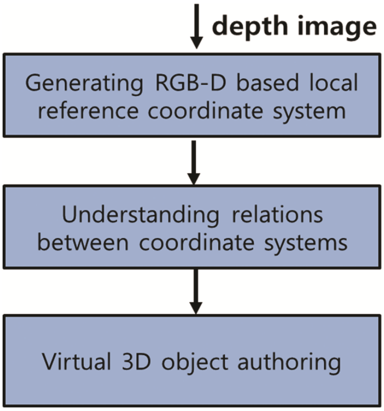
depth image
Generating RGB-D based local reference coordinate system
Understanding relations between coordinate systems
Virtual 3D object authoring

Fig. 2. Pipeline of the AR authoring system

Figure 3 shows the concept of our AR authoring system. Before working remote collaboration system, we organize a shared AR working space where the local and remote users perform a target task.

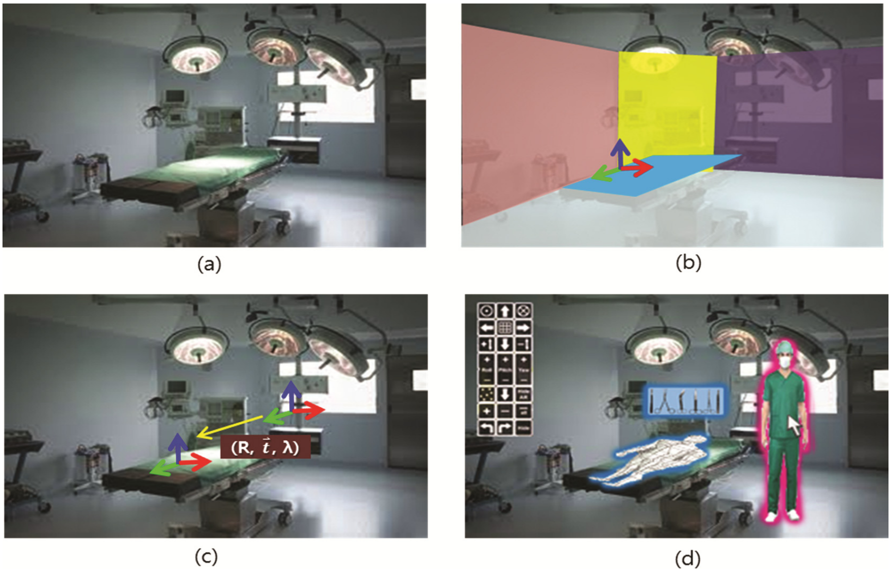

Fig. 3. The concept of AR authoring system: (a) real space, (b) local reference coordinate is calculated based on plane, (c) SLAM based coordinates is converted local reference coordinate by translation matrix, (d) virtual contents are augmented in AR space with local reference coordinate system.

2.2.1 Local Reference Coordinate System

To generate local reference coordinate system, we first select an original point and find their rotation coordinate system by analyzing 3D point clouds acquired from RGB-D camera. Then, a user choose regions of interest (RoI) in a scene using mobile input device. The RoI is detected a circular with a radius of 50 pixels. After a selecting RoI, we estimate the planes using point clouds of RoI. The plane of parameters π_i (a,b,c,d) are estimated by RANSAC method [14] as follows:

$$\pi_i = \underset{a,b,c,d}{\operatorname{argmin}} \sum\nolimits_l^N \frac{|ax + by_l + cz_l + d|}{\sqrt{a^2 + b^2 + c^2}}, (x_l, y_l, z_l) \in RoI. \tag{1}$$

We assume that the maximum number of planes of RoI is three. It is possible that finding the original point of local reference coordinate system is expressed as a linear least squares problem.

$$p_{center} = \underset{x_i, y_i, z_i}{\operatorname{argmin}} \left\| \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} - \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} \right\|^2, \tag{2}$$

The point which has the minimum sum of squares of distance among planes will be selected as the original point of local reference coordinate system. The directions of three axis are expressed with three intersection lines on planes.

2.2.2 Adjusting SLAM Coordinates to Local Coordinates

Based on [15], we estimate a camera pose and build a 3D point map in an unknown scene. It is necessary to adjust SLAM-based coordinates system to local reference coordinate systems for seamless registration with virtual object in real space. For this, we calculate two relations for adjustment process. First relation is a scale unit. The scale parameter of SLAM-based coordinates system is randomly selected in the initialization stage. This scale parameter should be replacement to a real scale unit. Without refining the scale parameter, the users couldn't register virtual object to the position they want. The scale ratio parameter λ is calculated using distance from camera position to starting point of each coordinate systems. The depth of RGB-D camera is presented with a meter scale unit.

$$\lambda = \frac{\textit{distance from camera to origin in virtual scale units}}{\textit{distance from camera to origin in real scale units}}. \tag{3}$$

Second relation is translation matrix $P_{local,n}$ which transforms the points of SLAM coordinates to local coordinates at n_{th} frame. We first calculate the initial matrix R that represents transform between coordinates, and then we compute the motion matrix M_n for each frame. This matrix represents an accumulated camera motion from the initial frame to the current frame.

$$M_n = M_{n-1} \times \ldots \times M_1 \times M_0. \tag{4}$$

$P_{SLAM,n}$ which is the matrix of transforming points from world coordinates to SLAM based coordinates is computed by motion matrix and $P_{SLAM,0}$.

$$P_{SLAM,n} = M_{n-1} \times P_{SLAM,0}. \tag{5}$$

Matrix $P_{local,n}$ can be expressed by the following equation:

$$P_{local,n} = \frac{1}{\lambda} \times M_{n-1} \times \mathrm{R} \times M_{n-1}^{-1} \times P_{SLAM,n}, n \neq 0. \tag{6}$$

If we obtain this initial coordinates, we can apply relation matrix R, motion matrix M and the scale unit parameter to it. The translation matrix $P_{local,n}$ provides an augmented space to matching real space.

2.2.3 3D Contents Authoring

The shared common working space is built by smartphone gestures such as tap, pinch, and rotate. The smartphone device is better than user's bare hand as the input device, because it enables delicate arrangement of virtual object in real space.

2.3 Hands-Augmented Object Interaction

Tracking two hands is important for natural interaction with virtual objects. There are two main approaches for this. [16] utilizes a generative method to track full articulations of two hands. The generative method has an advantage with respect to good generalization and continuous solution. However, it has a weakness that the solution falls easily into local minima if the solution is not good in previous frame. [17] utilizes a discriminative method to detect full articulations of two hands. This method has an advantage that the solution in the present frame is not affected by the solution of previous frame. So, it can detect full articulation of two hands in single frame. However, it gives a discrete solution and tends to the overfitting on training data. To complement the weakness of each method, our method utilizes both of the generative and discriminative methods.

2.3.1 Hand Feature Extraction

The proposed method utilize a convolutional neural network (CNN) with heterogeneous input devices for hand-virtual object interaction which illustrated in Fig. 4. As the used input devices, we are a RGB-D camera and an IMU sensor. The hand image is parsed into a normal deep network with convolution and pooling layer. The activation function is used as a rectified linear unit. The feature map obtained from last pooling layer is unified with 3 DoFs data from the IMU sensor. The remaining layers is fully connected layer so that we can unify the two heterogeneous data. Consequently, we get some heat maps that detect the position of joints with the highest probability.

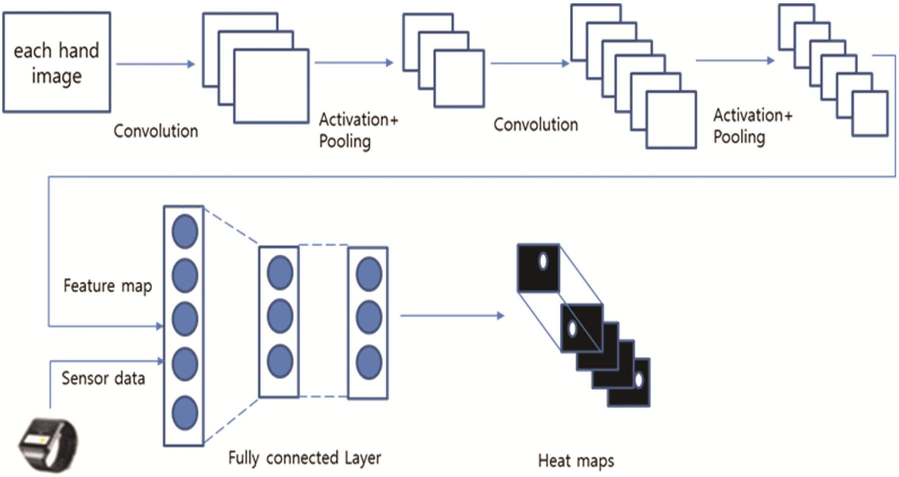

Fig. 4. The proposed CNN with heterogeneous inputs

2.3.2 Hand Pose Estimation

To estimate full articulations of the hands, we adopt two optimization schemes (See Fig. 5). The input datum are the segmented hand images and the heat maps generated from the proposed feature extraction algorithm. First, the inverse kinematics (IK) optimizer is conducted. This algorithm has good advantage about fast convergence. The designed objective function Eq. (7) calculates the error between fingertip and target position so that it find the parameter of articulations. The J is Jacobian matrix and $\vec{e}$ is a vector from source to target, $\Delta\theta$ is the variance of joint parameter. However, if the feature extraction is not accurate in some case, IK algorithm would fail.

$$\mathrm{E}_1 = \left\| J\Delta\theta - \vec{e} \right\|^2 + \lambda \|\Delta\theta\|^2. \tag{7}$$

To overcome this problem, the particle swarm optimization (PSO) method is employed. This is conducted only when the solution is not satisfied by a threshold. The objective function Eq. (8) is to measure the discrepancy between observation and hand model. O_i is 3D point in observation and M_i is 3D point in model. $w_{i,j}$ is the weight between model and observation.

$$\mathrm{E}_2 = \sum_i \sum_j w_{i,j} \left\| O_i - M_j \right\|_2 \tag{8}$$

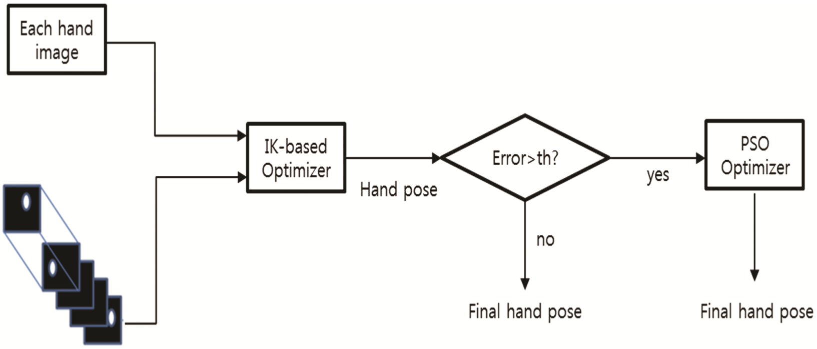

Fig. 5. The process of hand pose estimation

3 Implementation

3.1 Hardware Configuration

We configure our prototype system using commodity devices. Our system consists of a computing unit (PC) for computation, a video see-through HMD (HMD and stereoscopic RGB camera) for visualization, a near-range depth sensor and a smartwatch for bimanual hand tracking, a smartphone for AR authoring. We additionally use exocentric body tracker for body tracking.

For a video see-through HMD, we use Oculus Rift DK2 and attach Ovrvision stereoscopic RGB camera. Oculus Rift DK2 supports position and rotation tracking by external HMD tracker. For a near-range depth sensor, we use a Creative Senz3D. We use a Samsung Gear Live for smartwatch. Finally, we used a Microsoft Kinect v2 for body tracker.

3.2 Software Configuration

We implement the initial prototype in Unity Engine [18]. Figure 6 illustrates the components and their relationship of proposed system.

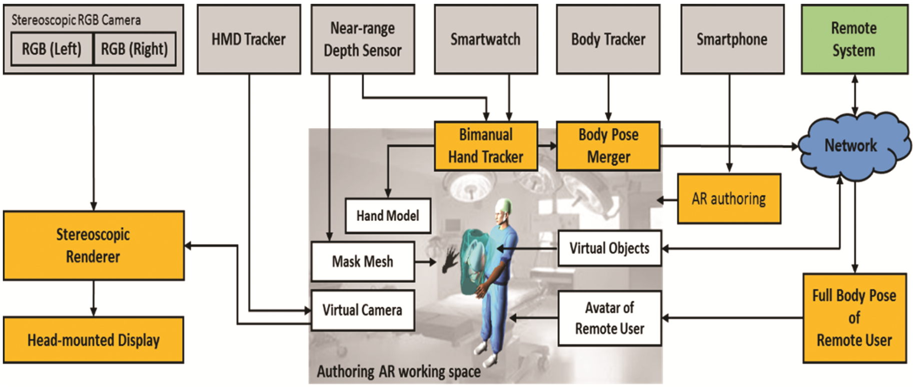

Fig. 6. Detail diagram of proposed framework

For interactive AR contents authoring, we use the positions, rotations and touch directions information which comes from a smartphone input device. With these information, a user can organize his/her AR working space and share the space with remote users.

For hands-augmented object interaction and collaborations with remote users, we utilize a near-range depth sensor and a smartwatch for bimanual hand tracking. Bimanual hand tracking result is used for virtual object manipulation. Also, user's bimanual hand posture information is combined with body pose information from body tracker, and generates combined body-hand pose information is sent to the remote space through network in real-time. At the same time, remote user's combined body and hand pose information is received in real-time, and is used for manipulating avatar movement.

We also utilize point cloud from a near-range depth sensor, to generate occlusion mask mesh which is used for enhancing user's depth perception between hand and virtual object. Then, final virtual scene is merged with real world view, and HMD displays virtual-real combined image.

3.3 Initial Result

Figure 7 shows initial result of our prototype of HMD-based collaboration system. Remote user is summoned to local user's space as a virtual avatar, and both users use bimanual hand gesture to interact with virtual objects. Unlike previous collaboration system [3], our system supports two hands interaction with virtual objects by tracking their hands. Furthermore, after integration with AR authoring method, our system can enable a local user to organize a user-friendly working space without any professional programming skills.

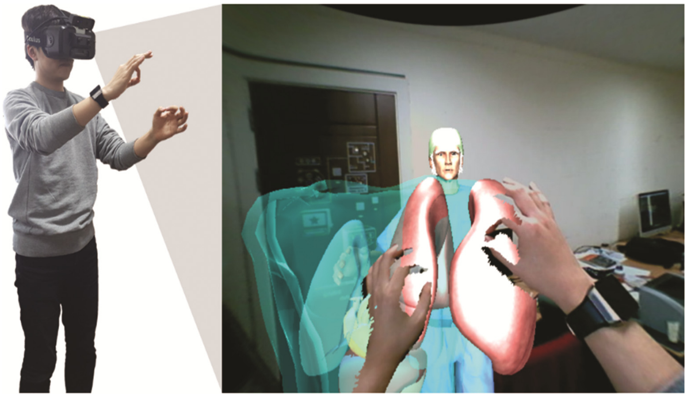

Fig. 7. Initial results: user uses bimanual hand gesture to interact with virtual objects

4 Conclusions and Future Works

In this paper we have presented a novel unified framework for HMD-based remote collaboration using interactive AR authoring and two hands tracking, which enables a local user to organize a user-friendly working space without any professional programming skills, and collaborate physically remote users through intuitive hands-augmented object interactions. Preliminary implementation result shows its strong possibility as a future remote collaboration platform. We expect that the proposed framework can be applicable to many AR collaborative applications such as urban planning, games, medical surgery education, and so on.

As the future works, we plan to the development of two hands tracking with wearable sensor and integration AR authoring and collaboration system.

Acknowledgements. This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2015M3A6A3073 746, NRF-2014R1A2A2A01003005).

References

1. Azuma, R.: A survey of augmented reality. Presence **6**(4), 355–385 (1997)
2. Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, B.: Recent advances in augmented reality. IEEE Comput. Graph. Appl. **21**(6), 34–47 (2001)
3. Noh, S., Yeo, H., Woo, W.: An HMD-based mixed reality system for avatar-mediated remote collaboration with bare-hand interaction. In: Eurographics-ICAT-EGVE (2015)
4. Jo, D., Kim, K., Kim, G.: SpaceTime: adaptive control of the teleported avatar for improved AR tele-conference experience. CAVW **26**, 259–269 (2015)
5. Ha, T., Billinghurst, M., Woo, W.: An interactive 3D movement path manipulation method in an augmented reality environment. Elsevier Interact. Comput. **24**(1), 10–24 (2012)
6. Ha, T., Feiner, S., Woo, W.: WeARHand: head-worn, RGB-D camera-based, bare-hand user interface with visually enhanced depth perception. In: IEEE ISMAR, pp. 219–228 (2014)

7. Jang, Y., Noh, S., Chang, H., Kim, T., Woo, W.: 3D finger CAPE: clicking action and position estimation under self-occlusions in egocentric viewpoint. IEEE Trans. Vis. Comput. Graph. **21**(4), 501–510 (2015)
8. Wang, Y., Langlotz, T., Billinghurst, M., Bell, T.: An authoring tool for mobile phone AR environments. In: Proceedings of New Zealand Computer Science Research Student Conference, pp. 1–4 (2009)
9. Higuchi, K., Chen, Y., Chou, P., Zhang, Z., Liu Z.: ImmerseBoard: immersive telepresence experience using a digital whiteboard. In: CHI (2015)
10. Beck, S., Kunert, A., Kulik, A., Froehlich, B.: Immersive group-to-group telepresence. IEEE Trans. Vis. Comput. Graph. **19**(4), 616–625 (2013)
11. Jung, J., Hong, J., Park, S., Yang, H.: Smartphone as an augmented reality authoring tool via multi-touch based 3D interaction method. In: Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry, pp. 17–20 (2012)
12. Langlotz, T., Mooslechner, S., Zollmann, S., Degendorfer, C., Reitmayr, G., Schmalstieg, D.: Sketching up the world: in situ authoring for mobile augmented reality. Pers. Ubiquit. Comput. **16**(6), 623–630 (2012)
13. https://www.google.com/atap/project-tango/
14. Fischler, M.A., Robert, C.B.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
15. Klein, G., David, M.: Parallel tracking and mapping for small AR workspaces. In: ISMAR (2007)
16. Oikonomidis, I., Lourakis, M., Argyros, A.: Evolutionary quasi-random search for hand articulations tracking. In: CVPR (2014)
17. Rogez, G., Khademi, M., Supančič III, J.S., Montiel, J.M.M., Ramanan, D.: 3D hand pose detection in egocentric RGB-D images. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8925, pp. 356–371. Springer, Heidelberg (2015)
18. http://unity3d.com/