Robust Hybrid Tracking with Life-Size Avatar in Mixed Reality Environment

Q.C.T. Tran, S.P. Lee, W.R. Pensyl, and D. Jernigan

Interaction & Entertainment Research Centre – IERC
Nanyang Technological University, Singapore
{qui_tran,splee,wrpensyl,DJernigan}@ntu.edu.sg

Abstract. We have developed a system which enables us to track participant-observers accurately in a large area for the purpose of immersing them in a mixed reality environment. This system is robust even under uncompromising lighting conditions. Accurate tracking of the observer's spatial and orientation point of view is achieved by using hybrid inertial sensors and computer vision techniques. We demonstrate our results by presenting life-size, animated human avatars sitting in real chairs, in a stable and low-jitter manner. The system installation allows the observers to freely walk around and navigate themselves in the environment even while still being able to see the avatars from various angles. The project installation provides and exciting way for cultural and historical narratives to be presented vividly in the real present world.

1 Objective and Significance

The objective of this project is to create a low-cost, easy-to-set-up, robust and reliable, interactive and immersive augmented and mixed reality system which presents life-size, animated human avatar in a cultural heritage environment.

The significance of this project is that it provides an environment where people can interact with historical human characters in a cultural setting. Other than for cultural and historical installation, this work has the potential of being further developed for the purpose of education, art, entertainment and tourism promotion.

2 Introduction

The motivation for developing this project is to allow people to experience pseudo-historical events impressed over present day real world environment (we have set our project at the famous Long Bar at the Raffles Hotel in Singapore). It is an augmented reality installation which involves re-enactment of the famous people who frequented the bar in the early 20th century. It uses augmented reality technology to develop both historical and legendary culturally significant events into interactive mixed reality experiences. Participants wearing head-mounted display systems witness virtual character versions of various notable figures, including Somerset Maugham, Joseph Conrad, and Jean Harlow, immersed within a real world environment modeled on the Raffles Hotel Long Bar they had frequented. Through the application of research in

R. Shumaker (Ed.): Virtual and Mixed Reality, LNCS 5622, pp. 503–510, 2009.

tracking, occlusion, and by embedding large mesh animated characters, this installation demonstrates the results of the technical research and the conceptual development and presentation in the installation. Moreover, requiring that our work eventually be located in the Long Bar provides us with the motivation to create a system that can accommodate compromising lighting conditions and large open spaces.

Tracking of human subjects in a large area indoor mixed reality application has always been a challenge. One of the most crucial parts of an appealing mixed reality presentation is the smooth, jitterless and accurate rendering of 3D virtual objects onto the physical real world. This involves primarily the tracking of user's point of view which in most cases are the spatial and orientation information of the viewing camera attached to user's head. Conventionally, fiducial markers are used in conjunction with ARToolkit [1] or MXRToolkit [2] in well-prepared, well-lit environment; however the use of markers poses several problems: they are not robust in a poor lighting environment; and placing and scattering markers in the environment makes the real world scene unaesthetic.

Previous work [3] tracked humans in an environment "whose only requirements are good, constant lighting and an unmoving background". In [4], capturing the human body requires a green recording room with consistent lighting. Our system, on the other hand, works in most unexpected, varying, artificial and/or ambient lighting condition.

The work in [5] provided excellent accounts and experimental results of hybrid inertial and vision tracking for augmented reality registration. The sensitivities of orientation tracking error were quantitatively analyzed, natural feature tracking and motion-based registration method that automatically computed the orientation transformation (among different coordinate systems) was presented. Our work is different in that the tracking cameras tracks only point light source, and they are statically mounted on the wall instead of attached to the observer/user's head.

3 Overview

This work was intended to be set up in the famous and historical Long Bar in Singapore. People wearing HMDs will be able to see virtual avatar of famous historical human character talking and interacting with them.

Our work uses a hybrid approach – tracking an active marker while at the same time tracking the movement of the observer's (camera's) frame with the use of inertial sensors. The active marker is made up of an infrared (IR) light-emitting diode (LED) mounted on the user's HMD. In our system, to detect IR LED, instead of using normal cameras, we use Nintendo Wii Remotes as vision tracking devices. This low-cost device can detect IR sources at up to 100 Hz, which is very suitable for real-time interaction systems. Furthermore, an inertial sensor consisting of 6DOF is attached to the HMD to detect the rotation and movement of user viewpoints.

There are two ways in which the user can interact with the virtual character. Upon request by the virtual character, the user could fill the physical glass with wine by first picking the glass up from the table. He will then "fill" the glass with "wine", and put the glass on anywhere on the table. The virtual character will then pick up the glass from that location. Another interaction is that when the virtual character asks the user

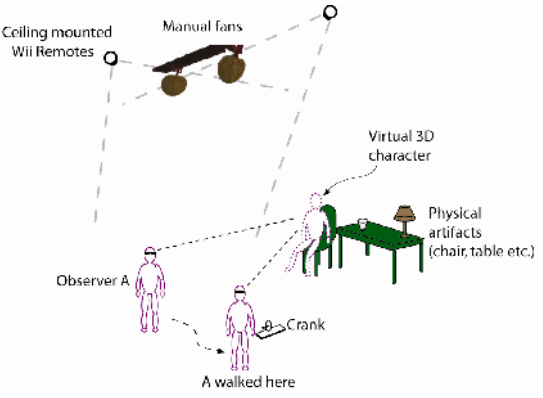

Fig. 1. Setup of the large area robust hybrid tracking system. The chair, table, lamp and glass are physical objects in the real world. The manual fan on the ceiling is virtual; it is controlled by a real physical crank. The "human" in dashed-line and sitting in the chair is a 3D avatar.

to crank up the virtual fans (which were installed on the ceiling) when it gets hot. The user then cranks up a physical device to control the speed of the fan. If the fans are too fast the virtual character will ask the user to slow down, and vice versa.

Figure 1 shows the system setup. The observer wears a HMD and can move freely around within the designated area. There is an empty chair, a table and other physical artifacts (glass, etc) in the demonstration area. As the observer looks at the chair, he sees a life size, 3D animated human character sitting in the chair.

4 System Details

4.1 System Block Diagram

Figure 2. shows the block diagrams of the system. The peripheral hardware are shaded in blue.

The details of each part will be described in the following subsections.

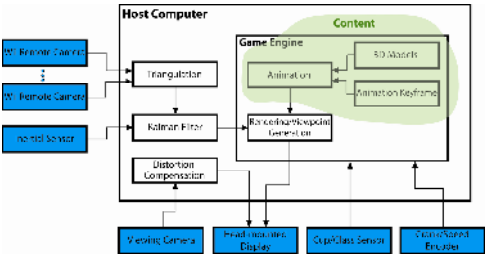

Fig. 2. Block diagram of the system

4.2 Tracking

The tracking of user's point of view through the HMD is essentially the continuous tracking of the camera (mounted on the HMD) frame's spatial coordinate and orientation, with respect to the world frame of reference set in the real environment. It is a combination of computer vision (camera tracking) and inertia sensing.

Low-cost Vision-Based Tracking of Head Position. Conventional ARToolkit or MXRToolkit markers are not suitable for vision-based tracking in large area, uncompromising lighting environments. Jittering and loss of tracking due to the lighting conditions seriously hampers accurate tracking, adversely impacting the audiences' aesthetic experience. In view of this, we have devised an active beacon by using IR LED as a position tracking device. Instead of employing black and white markers, the observer wears light-weight HMD which has an IR LED attached to it.

We use two Nintendo Wii Remotes as vision tracking devices. The Wii consists of a monochrome camera with resolution of 128x96 pixels, with an IR-pass filter in front of it. The Wii Remote cameras are installed in the ceiling to track the location of the observer. This tracking provides positional information only and does not provide the orientation information of the head.

The advantages of using Wii Remote cameras are manifold: low-cost, easy setup, high "frame rate" (in fact only processed images – the coordinates of the tracked points – are sent to the host) and wireless. The host computer is relieved from processing the raw image; and an optimal triangulation technique is all that is needed to obtain the depth information of the IR LED.

Inertial Sensor-based Tracking of Head Orientation. The inertial sensor is attached to the HMD, above the viewing camera. It provides drift-free 3D acceleration, 3D rate gyro and 3D earth-magnetic field data [6]. With sensor fusion algorithm, the 3D orientation (roll, pitch and yaw) data of the sensor's coordinate frame can be found.

As we are tracking different sources (IR LED for position and inertial sensor for rotation), the calibration process which is already important becomes even more critical. In the next part, we will describe the steps we have done in our calibration process.

4.3 Calibration

Intrinsic Calibration for Cameras. We used Camera Calibration Toolbox for Matlab [7] to calculate the intrinsic parameters of our viewing camera. The result we got from the Toolbox is an intrinsic matrix I:

$$I = \begin{bmatrix} f_x & S & O_x \\ 0 & f_y & O_y \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

Where: f_x , f_y : the focal length expressed in units of horizontal and vertical pixels.

O_x , O_y : the principal point coordinates.
S: the skew coefficient defining the angle between the x and y pixel axes.

From these parameters, the projection matrix that will be used in the game engine is calculated as follow: (in the same way as in ARToolkit [1])

$$P = \begin{bmatrix} \frac{2\times f_x}{Width} & \frac{2\times S}{Width} & \frac{2\times O_x}{Width}-1 & 0 \\ 0 & \frac{2\times f_y}{Height} & \frac{2\times O_y}{Height}-1 & 0 \\ 0 & 0 & \frac{(gFar+gNear)}{(gFar-gNear)} & \frac{-2\times gFar\times gNear}{(gFar-gNear)} \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (2)$$

Where: *Width, Height*: the resolution of the camera
gFar, gNear: the far and near clipping planes

Similar to the viewing camera, the intrinsic parameters of the cameras inside wiimotes also need to be calculated. To do that, we use a small board with 4 IR LEDs. Wiimotes track these 4 IR sources simultaneously while the board is moved around. A few frames are captured and intrinsic parameters can be calculated.

Extrinsic Calibration. In this step, the transformation from the coordinate system of sensors to the coordinate system of the viewing camera needs to be calculated. To do that, we are using the method similar to the method described in [8].

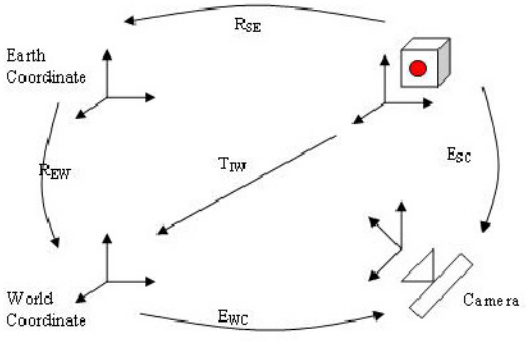

Fig. 3. The transformations between all the coordinate system

The relations between all the coordinate system are illustrated in Figure 3. In the figure, we use the following annotation: ***R*** for rotation, ***T*** for translation, ***E*** for transformation (E = [R|T]). $\boldsymbol{R}_{SE}$ is the rotation read from the inertial sensor, which is aligned to the Earth coordinate system. $\boldsymbol{T}_{IW}$ is the translation from the IR Led to the world coordinate system. T_{IW} is calculated by using the 2 wiimotes. E_{SC} and R_{EW} are the 2 unknown parameters that need to be calculated. As can be seen in Figure 3:

$$
\begin{aligned}
& E_{SC} = E_{WC}\left[R_{Ew}R_{SE} \mid T_{IW}\right] \\
& \Rightarrow E_{SC} = [R_{WC}T_{WC}]\left[R_{Ew}R_{SE} \mid T_{IW}\right] \\
& \Rightarrow \begin{cases} R_{SC} = R_{WC}R_{Ew}R_{SE} \\ T_{SC} = R_{WC}T_{IW} + T_{WC} \end{cases}
\end{aligned} \tag{3}
$$

During the calibration process, we put a chess board at the center of the world coordinate system to compute E_{WC}. At the same time, T_{IW} and R_{SE} are also captured. After capturing enough combinations (E_{wc}, $T_{IW,}$ R_{SE}), R_{SC} and T_{SC} can be calculated in the form of solving the "AX=XB" equation system [8]. The same for R_{EW}.

4.4 Interaction

We provide two interesting scenarios in which user can interact with the 3D virtual human avatar. The first scenario is when the virtual human asks for his glass to be filled up with wine. The user himself, or another user (the bartender, as shown in Figure 4) will then pick up the physical glass from the table, fill it up, and put in on an arbitrary place on the table. The virtual human is supposed to pick up the glass from that location. A second scenario is that the user is supposed to respond to the avatar request of fanning the virtual ceiling fan. The details of the interaction and sensing techniques are described in the following sub-sections.

Arbitrary Placement of the Cup/Glass. Ferromagnetic metal detection sensors are installed at various spots under the table. A small, thin piece of iron is attached to the underneath of the glass. As user/bartender puts down the glass onto an arbitrary spot on the table, its position will be sensed and sent to the host computer. The virtual human avatar then performs a pre-animated action to pick up the virtual glass (Figure 4).

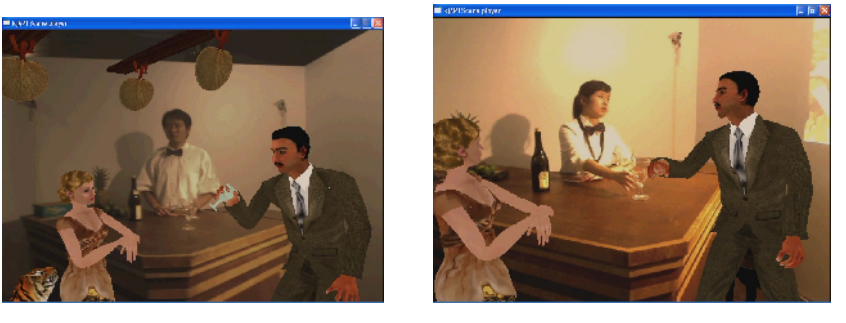

Fig. 4. First person view through the HMD. On the left figure, the virtual human pick up the "glass" (virtual) after the bartender put the glass (real) on the table. On the right, the virtual human puts the "glass" down and the bartender picks it up. Notice the virtual fans on the ceiling.

Cranking the Virtual Fan. A physical crank (Figure 5) is provided as a device to control the virtual ceiling fans. A rotary encoder is embedded inside the crank and it is used to measure the speed at which the user rotates the crank. Its rotational speed corresponds to that of the virtual ceiling fan. Notice in Figure 4 the virtual fans on the ceiling. The virtual human will react to the fan speed, such as complaining that the weather is hot and the fan is too slow.

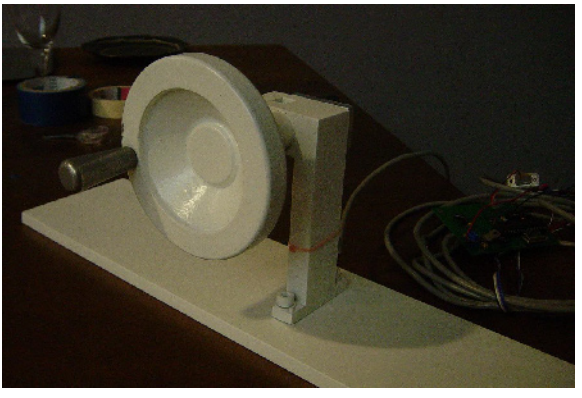

Fig. 5. The rotary crank used to control the virtual fan. It is connected to the host computer via a serial cable.

5 Conclusion and Future Work

This work provides users with enhanced experience in a large area immersive mixed reality world. The robust and stable rendering of the 3D avatars is made possible by accurate tracking which involves hybrid computer vision and inertial sensors. Users are able to interact with the avatars with the aids of various sensing techniques.

This work is originally intended for use in a cultural and heritage setting due to its large space and live-size avatars nature. However it can also be extended to applications such as online role-playing game, 3D social networking and military simulation.

One disadvantage of the hybrid tracking method is that inertial sensor and LEDs have to be mounted onto the HMD; this essentially makes it cumbersome and uncomfortable to wear. We propose to use, in next stage of development, natural feature tracking technique in place of the hybrid tracking. One such technique is parallel tracking and mapping (PTAM) [9].

Currently the position of the glass is discrete, since only a few ferromagnetic sensors are place underneath the table. Future version of this glass tracking technique would be a “continuous” and high resolution one; possible technologies which could be used are capacitive sensing and electric field sensing (similar to the tablet PC technology). Pre-animated pick-up-the-glass motion could not possibly be used because theoretically there are infinite picking up motions from infinite number of spots on the table. We propose that inverse kinematics be incorporated in the motion of the virtual human avatar when he picks up the glass from the table.

Physics would also be incorporated within the interaction scenarios to make it more realistic. For example, too strong the speed of the fan would blow the virtual human's hair and mess it up.

Haptics could also be used in this work. The user would be able to pat the virtual human's shoulder, or even shake hands with him.

References

1. ARToolkit, http://www.hitl.washington.edu/artoolkit
2. MXRToolkit, http://sourceforge.net/projects/mxrtoolkit
3. Sparacino, F., Wren, C., Davenport, G., Pentland, A.: Augmented Performance in Dance and Theatre. In: International Dance and Technology, ASU, Tempe, Arizona (1999)
4. Nguyen, T.H.D., Qui, T.C.T., Xu, K., Cheok, A.D., Teo, S.L., Zhou, Z.Y., Allawaarachchi, A., Lee, S.P., Liu, W., Teo, H.S., Thang, L.N., Li, Y., Kato, H.: Real Time 3D Human Capture System for Mixed-Reality Art and Entertainment. IEEE Transaction on Visualization and Computer Graphics (TVCG) 11(6), 706–721 (2005)
5. You, S., Neumann, U., Azuma, R.: Hybrid Inertial and Vision Tracking for Augmented Reality Registration. In: Proceedings of the IEEE Virtual Reality, Washington, DC (1999)
6. XSens: MTx 3DOF Orientation Tracker, http://www.xsens.com/Static/Documents/UserUpload/dl_42_leaflet_mtx.pdf
7. Camera Calibration Toolbox for Matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/
8. Baillot, Y., Julier, S.J.: A Tracker Alignment Framework for Augmented Reality. In: Proceedings of 2nd IEEE and ACM International Symposium of Mixed and Augmented Reality (2003)
9. Klein, G., Murray, D.: Parallel Tracking and Mapping for Small AR Workspaces. In: Proceedings of 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, ISMAR 2007 (2007)