Occupant Location and Gesture Estimation in Large-Scale Immersive Spaces

Devavrat Jivani
Rensselaer Polytechnic Institute
110 8th Street, Troy, NY 12180
jivand@rpi.edu

Gyanendra Sharma
Rensselaer Polytechnic Institute
110 8th Street, Troy, NY 12180
sharmg3@rpi.edu

Richard J. Radke
Rensselaer Polytechnic Institute
110 8th Street, Troy, NY 12180
rjradke@ecse.rpi.edu

The research in this paper was supported by Rensselaer Polytechnic Institute and by IBM Research via Rensselaer's Cognitive and Immersive Systems Laboratory.
Living Labs: Measuring Human Experience in the Built Environment, in association with CHI'18., April 22, 2018, Montréal, Canada.

Abstract
Large-scale immersive environments can convey a large amount of information to many participants simultaneously. When equipped with occupant-awareness, not only does the ability of the participants to interact with the environment increase dramatically, but also the environment's understanding of its audience. Using ceiling-mounted time-of-flight distance sensors, we discuss how person tracking, body orientation estimation, and pointing gesture recognition can augment an immersive environment to become simultaneously aware of all its participants. The testbed for our exploration is a large circular space 12 m in diameter enclosed by a 360 degree display 4.3 meters high, which can comfortably house more than 30 people at a time. This large-scale occupant-aware immersive environment is an ideal space to simulate different "living lab" scenarios or conduct group behavioral studies.

Author Keywords
Immersive environments, mid-air pointing, person tracking, orientation estimation, multi-user spaces

ACM Classification Keywords
H.5.m. [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces; Interaction styles []

Introduction

Immersive, interactive, and intelligent environments have great potential to augment group intelligence and aid in decision making. For instance, a hospital's tumor board convening to discuss the treatment plan for a particular patient could use a large display to project test results and imaging scans [2]. Furthermore, an environment with detailed awareness of its participants' location, orientation, and gestures could actively allow local interaction with displayed information, present targeted information to specific users, and cue users' attention to different areas of the room.

Our work explores the visual intelligence aspect of immersive environments, using information obtained from distance sensors that unobtrusively observe the entire space and measure its interior as a dynamic 3D point cloud. In particular, we conduct studies in a large-scale, circular environment with a 360 degree screen in the Experimental Media and Performing Arts Center (EMPAC) at Rensselaer Polytechnic Institute (Figure 1). The screen is 12 m in diameter, 37 m in circumference, and 4.3 m high, allowing for a truly immersive experience that can simultaneously accommodate more than 30 people. We instrumented this space with a network of ceiling-mounted, downward-pointed Microsoft Kinect sensors, with custom algorithms to estimate the location, body pose, and pointing gestures of the participants. We discuss the challenges and solutions of working in this large-scale environment, including poor illumination and the need for accurate calibration between multiple sensors.

Related Work

Large-scale immersive systems and interaction techniques are active areas of research and experimentation. Designing interaction techniques for large display systems that go well beyond traditional devices like the mouse and keyboard

Figure 1: The circular immersive environment at EMPAC, RPI.

is critical to reach the full potential of such environments. In particular, markerless body tracking and orientation is a key issue for immersive and virtual environments [8].

Interaction designs greatly benefit from spatial awareness of the location and orientation of participants, especially in the context of large screen displays. Ballendat et al. [1] discussed how various implicit and explicit interaction techniques can be regulated based on proxemic information obtained through sensing technologies. In terms of the underlying technology, depth cameras have been used previously for human tracking in indoor spaces [7, 10, 11]. Beyond tracking, the Kinect's built-in skeletal tracking allows the easy detection of gestures like pointing, leveraging the automatic estimation of joints in the human body [4, 5, 6].

Building a Unified Point Cloud

Because most of the illumination in a large-screen environment is provided by the screens themselves, it is generally dimly lit (see Figure 1). We thus require a tracking system that does not rely on visible light like RGB cameras do. Mi-

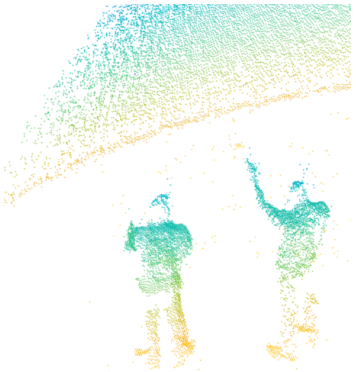

Figure 2: Close-up 3D point cloud view of people using the space

crosoft Kinects, which include depth sensors operating on the time-of-flight principle, have the distinct advantage of providing accurate 3D information about the scene using an infrared transmitter and high-resolution low-illumination imaging sensor. As shown in Figures 2 and 3, the resulting depth maps provide enough 3D information to estimate the location, pose, and gestures of the participants without any additional markers, body-mounted sensors, or hand-held devices. The participants are able to explore and inhabit the space unencumbered by any additional equipment.

To cover the 12 m diameter circular space, we use 6 Kinect sensors mounted to a ceiling grid at a height of 5.5m from the ground facing downwards. The Kinects are mounted on the ceiling rather than being placed at ground level to avoid obstructing the screen and breaking the immersion provided by the environment. However, the overhead configuration of the Kinects means that automatically estimated skeletal tracking information is no longer available using the Microsoft Kinect SDK. Instead, we treat the Kinects as distance sensors that relay 3D point cloud information from their individual fields of view (which allows for more privacy compared to visible-light cameras).

Using open-source drivers [9] we read the data from all six Kinects simultaneously onto a single machine. To align all the Kinects against each other with respect to a global reference frame, we performed a one-time calibration process at the time of installation. This involved a 75cm$\times$100cm flat checkerboard pattern placed at various positions and orientations in the overlapping fields of view between each pair of neighboring Kinects. The sets of corresponding 3D checkerboard points are used to determine the 3D transformation between the Kinect coordinate systems using the iterative closest points algorithm [3]. Once the calibration is complete, we can construct a real-time 3D view of the entire

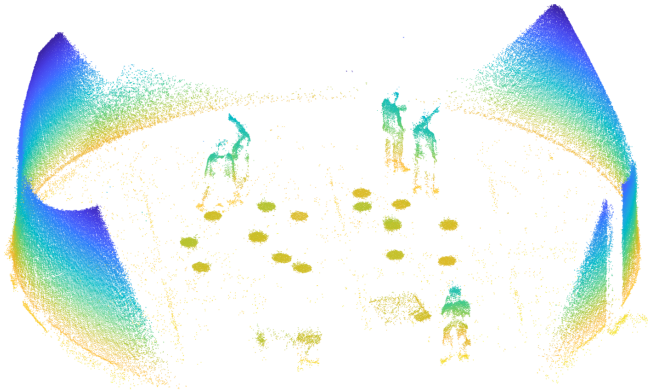

Figure 3: A unified point cloud of the space with several participants present, obtained after the multi-Kinect calibration process.

environment as shown in Figure 3.

Tracking Occupants

Person tracking is fundamental to allow the system to analyze the interactions between the screen and its potentially large number of users. It is important not only to detect people but also to assign and maintain unique tags for each participant, so that individual trajectories and activities can be recorded for analysis. Participants are detected from the view of each individual Kinect by performing intensity (height) thresholding and background subtraction in the depth map to eliminate the clutter of any static objects in the scene.

Following morphological operations on the background-subtracted image to help reduce the noise, we obtain in-

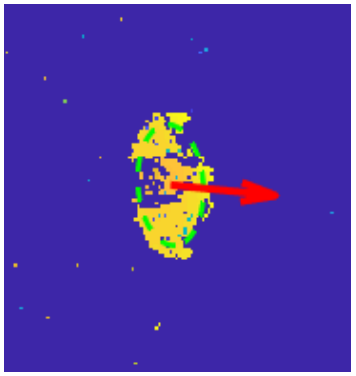

Figure 4: Each participant's orientation is estimated by fitting an ellipse (dotted green) to his/her head and shoulders. The arrow (red) indicates the orientation for a particular frame.

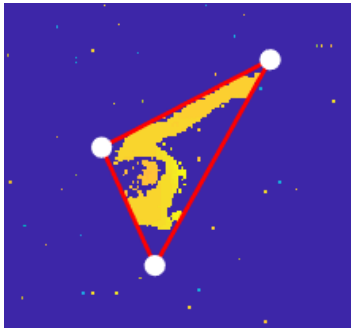

Figure 5: Triangle fitting to detect the coordinates of an outstretched hand to build a 3D pointing vector.

dividual pixel blobs representative of each person in the scene. A unique identifier is assigned to each person as they walk in from the single point of entry/egress in the space and maintained for the duration of the meeting. The calibrated coordinate system obtained previously ensures that the tag is maintained as a participant moves from the field of view of one Kinect into another. The location information can be used to enable localized interaction with the part of the screen in each participant's immediate vicinity, such as pointing.

Estimating Body Orientation

An unimpeded birds-eye view of the room allows us to go beyond tracking the participants to estimating their body orientation. In conjunction with location tracking information, this can be useful for enabling local interactions and for estimating user/group focus. For example, a pair of participants might be engaged in close conversation with each other, or a participant might be looking at a particular piece of information on the screen. In either case, accurate estimates of each participant's body orientation information are critical to allow the environment to take appropriate action.

An ellipse is fit to each participant's head and shoulder image, available from the person blobs detected at the tracking stage. By combining the minor axis of the ellipse with historical information about the participant's motion trajectory, we estimate the orientation angle for each participant in the range [0,360) degrees, as illustrated in Figure 4.

Interacting by Pointing

Conventional interaction techniques such as a mouse restrict the movements of the participants to a small area and defeat the purpose of a large immersive environment. We can avoid this pitfall by allowing gestural pointing input to manipulate different elements on the screen. Thus, we must be able to detect when a pointing gesture is being performed and extract relevant information. Here, we take an outstretched hand as indicative of a pointing gesture. This is easy in our system since the shape of the body contour from the overhead Kinect view during pointing is distinct from the shape when a person is standing upright with arms by his/her side. We detect the pointing event by setting a threshold on the ellipse size used to fit a person's body contour.

Once pointing is detected, we obtain a pointing vector determined by two 3D points on the head and the hand. When the Kinect 3D coordinate system is calibrated against the curved screen coordinates, this vector can be used to manipulate specific elements on the screen and provide visual feedback to the participant. We obtain the head point by finding the local height maximum in the head-shoulders region of a person depth map. We obtain the hand point by fitting a triangle to the outstretched body contour (Figure 5). The vertex of the triangle farthest from the head point is taken as belonging to the hand. A 3D pointing vector is obtained by connecting the 3D coordinates of a person's head to their hand (Figure 6).

Discussion and Conclusions

Indoor built environments such as offices, museums, and transit hubs will likely be equipped with large screens and some level of awareness about their occupants in the near future. Immersive occupant-aware environments of the sort described here have promising applications as testbeds to study large groups, especially in the context of interaction techniques and group dynamics.

Next steps in our particular project include the integration of visual tracking with acoustical tracking, using both lapel-worn microphones and room-mounted microphone arrays.

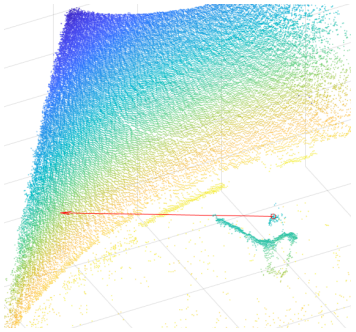

Figure 6: The 3D pointing vector connecting the top of the head to the hand determines the interaction point on the screen.

If the speech of each person can be clearly extracted, it can be processed using natural language understanding algorithms and passed into cognitive computing algorithms, enabling fine-grained, multi-modal occupant awareness. We hope to design a space in which the location, orientation, gesture, and speech of each participant can be analyzed in real time to enable group meetings in smart environments to become more engaging and productive.

REFERENCES

1. Till Ballendat, Nicolai Marquardt, and Saul Greenberg. 2010. Proxemic interaction: designing for a proximity and orientation-aware environment. In *ACM International Conference on Interactive Tabletops and Surfaces*. ACM, 121–130.
2. Jakob E Bardram, Thomas R Hansen, and Mads Soegaard. 2006. AwareMedia: a shared interactive display supporting social, temporal, and spatial awareness in surgery. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM, 109–118.
3. P. J. Besl and N. D. McKay. 1992. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14, 2 (1992), 239–256.
4. Harrison Cook, Quang Vinh Nguyen, Simeon Simoff, Tomas Trescak, and Dean Preston. 2015. A close-range gesture interaction with Kinect. In *Big Data Visual Analytics (BDVA), 2015*. IEEE, 1–8.
5. Hansol Kim, Yoonkyung Kim, Daejune Ko, Jinman Kim, and Eui Chul Lee. 2014b. Pointing gesture interface for large display environments based on the Kinect skeleton model. In *Future Information Technology*. Springer, 509–514.
6. Hansol Kim, Yoonkyung Kim, and Eui Chul Lee. 2014a. Method for user interface of large displays using arm pointing and finger counting gesture recognition. *The Scientific World Journal* 2014 (2014).
7. Evdokimos I Konstantinidis and Panagiotis D Bamidis. 2015. Density based clustering on indoor kinect location tracking: A new way to exploit active and healthy aging living lab datasets. In *Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on*. IEEE, 1–6.
8. Belinda Lange, A Rizzo, Chien-Yen Chang, Evan A Suma, and Mark Bolas. 2011. Markerless full body tracking: Depth-sensing technology within virtual environments. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.
9. Florian Echtler Lingzhu Xiang and others. 2016. libfreenect2: Release 0.2. (2016).
10. Muhamad Risqi Utama Saputra, Guntur Dharma Putra, Paulus Insap Santosa, and others. 2012. Indoor human tracking application using multiple depth-cameras. In *Advanced Computer Science and Information Systems (ICACSIS), 2012 International Conference on*. IEEE, 307–312.
11. Loïc Sevrin, Norbert Noury, Nacer Abouchi, Fabrice Jumel, Bertrand Massot, and Jacques Saraydaryan. 2015. Characterization of a multi-user indoor positioning system based on low cost depth vision (Kinect) for monitoring human activity in a smart home. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 5003–5007.