Multiscale Sensor Fusion for Display-Centered Head Tracking

Tianyu Wu* Benjamin Watson†

Visual Experience Lab, Computer Science
NC State University

Abstract

Emerging display usage scenarios require head tracking both at short (<1m) and modest (<3m) ranges. Yet it is difficult to find low-cost, unobtrusive tracking solutions that remain accurate across this range. By combining multiple head tracking solutions, we can mitigate the weaknesses of one solution with the strengths of another and improve head tracking overall. We built such a combination of two widely available and low-cost trackers, a Tobii Eye Tracker and a Kinect. The resulting system is more effective than Kinect at short range, and than the Tobii at a more distant range. In this paper, we discuss how we accomplish this sensor fusion and compare our combined system to an existing mechanical tracker to evaluate its accuracy across its combined range.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Virtual reality; Computing methodologies—Artificial intelligence—Computer vision problems—Tracking

1 Introduction and Related Work

Emerging multiview application scenarios are placing new demands on display and tracking technology. For low-latency remote conferencing with a standing desk, systems must prioritize what the viewer sees well, and eliminate visual artifacts at view boundaries [1]. For a small group pointing at a large display, systems must render only the views group members see. Such scenarios require both eye and head tracking at distances of 0.5 to 3 meters. Ideally, such systems would deliver an untethered and unobtrusive experience, which does not require users to don special tracking equipment.

The Tobii is best known for tracking eye saccades and fixations, but also reports head position. Unfortunately, at ranges beyond 1m, Tobii head tracking often fails: it is designed to operate within a 0.5 to 0.95 range [6]. On the other hand, Kinect v2 combines an RGB camera with an infrared camera, which gives it good depth accuracy. Kinect's near operating range is 0.5 meters, but its tracking accuracy declines below 0.8 meters range [3]. Although official support of Kinect v2 has ended, an active community maintains its drivers [2], and the next-gen Azure Kinect is now being sold. Neither Tobii nor Kinect functions reliably across the range required by our emerging display usage scenarios. We use a simple self-calibration to fuse these trackers into a single system that achieves this effective range.

Svoboda et al.'s prior work [5] is most similar to ours. Their method also eliminates hand-measurement of the relative positions of fused trackers with self-calibration. They place a well-known marker in view, and with that constraint, solve for a transformation that fuses the trackers. Our technique is similar, but simpler.

*e-mail: tianyu_wu@ncsu.edu
†e-mail: bwatson@ncsu.edu

2 Calibration Overview

The Tobii attaches to a display with a magnet, calibrates in less than a minute, and reports display-centered coordinates. The Kinect's coordinates are also centered around its hardware.

To fuse the Tobii and Kinect coordinate systems, we must define their spatial relationship. We could measure it by hand before tracking begins, which would make repositioning the Kinect (or the display and Tobii) problematic. Instead, we find the relationship automatically with self-calibration. Since both Tobii and Kinect report head position, we can use those coordinates as a constraint to solve for the relationship between the Tobii and Kinect coordinate systems. This technique may be easily extended to additional trackers as long as their tracking ranges overlap slightly; and to different types of trackers, as long as they are capable of identifying head position. We detail this self-calibration below.

3 Calibration Details

Our system consists of four components: one that retrieves data from the Tobii, the second from the Kinect, the third fuses these two data streams, and the fourth provides the fused data to applications. The first component reports tracking results in a Euclidean coordinate system centered around the display to which Tobii is attached. Our Kinect component also reports in a Euclidean system, but centered around the Kinect sensor. Our fused tracker can be in four states: only tracked by Tobii, only tracked by Kinect, tracked by both sensors, and not tracked. As soon as the user's head enters the range where it is tracked by both sensors we use their data to find the conversion from Kinect to Tobii.

Because both the Tobii and Kinect data are in Euclidean systems, an affine transformation T can transform between them. Data points p_{kinect} and p_{tobii} from the two sensors represent head position at the same time in the two coordinate systems. Given

$$T = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & x_0 \\ Q_{21} & Q_{22} & Q_{23} & y_0 \\ Q_{31} & Q_{32} & Q_{33} & z_0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, p_{kinect} = \begin{bmatrix} x_k \\ y_k \\ z_k \\ 1 \end{bmatrix} \text{ and } p_{tobii} = \begin{bmatrix} x_t \\ y_t \\ z_t \\ 1 \end{bmatrix},$$

we have $T \cdot p_{kinect} = p_{tobii}$, which can also be denoted as $p_{kinect}^{\top} \cdot T^{\top} = p_{tobii}^{\top}$. To solve for T, we need at least four distinct p_{kinect} and p_{tobii} pairs, at the bare (ideal) minimum. In practice, we compensate for sensor noise by using many more pairs, giving an overdetermined system $P_{kinect} \cdot T'^{\top} = P_{tobii}$, where

$$P_{kinect} = \begin{bmatrix} p_{kinect_1}^{\top} \\ p_{kinect_2}^{\top} \\ p_{kinect_3}^{\top} \\ \vdots \end{bmatrix}, P_{tobii} = \begin{bmatrix} p_{tobii_1}^{\top} \\ p_{tobii_2}^{\top} \\ p_{tobii_3}^{\top} \\ \vdots \end{bmatrix}, T' = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & x_0 \\ Q_{21} & Q_{22} & Q_{23} & y_0 \\ Q_{31} & Q_{32} & Q_{33} & z_0 \end{bmatrix}.$$

Instead of solving for an exact and most likely nonexistent T, we minimize $\|P_{kinect} \cdot T'^{\top} - P_{tobii}\|$ with a least squares solver. We use the resulting T' for Kinect-to-Tobii coordinate transformation.

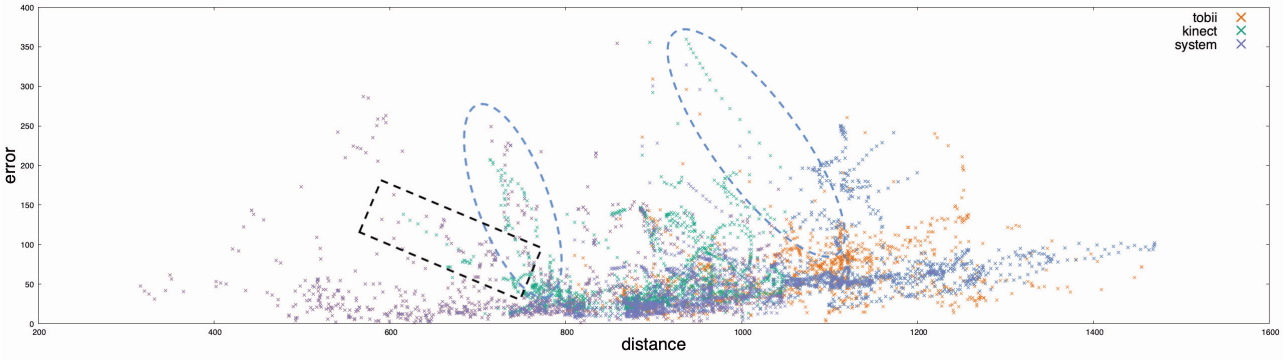

Figure 1: Error vs. distance (in mm) for the Tobii, Kinect and fused trackers. We removed three outlier positions for clarity.

Table 1: Root Mean Squared Error of Tobii, Kinect and fused positions vs. FARO over distance (in mm). Bold is minimum error.

Distance to Tobii	Tobii RMSE	Kinect RMSE	Fused RMSE
300 - 400	**45.74**	N/A	**45.74**
400 - 500	**77.40**	N/A	**77.40**
500 - 600	**95.17**	N/A	**95.17**
600 - 700	**50.32**	92.19	**50.32**
700 - 800	**67.00**	84.12	**67.00**
800 - 900	66.39	53.88	**53.00**
900 - 1000	57.01	93.91	**56.99**
1000 - 1100	**72.37**	84.29	76.30
1100 - 1200	**82.05**	100.06	100.06
1200 - 1300	95.81	**66.10**	**66.10**
1300 - 1400	**79.19**	81.83	81.83
1400 - 1500	**65.86**	95.14	95.14

4 Evaluation

To validate our system, we compared its output to a third position tracking device with reliable accuracy, the FARO mechanical arm, whose error was below 0.5 mm [4] — more than adequate for our evaluation. During setup, Kinect can be placed anywhere that overlaps its range well with Tobii's. We placed FARO to the side to reduce interference with Kinect. During testing, we asked the user to move widely within the FARO arm's limits at various speeds. We placed Kinect close to Tobii, so the two devices reported similar distances, otherwise the evaluation results might be quite different.

During evaluation, we found that using ~ 50 points to solve for T' minimized error; this number may vary with different numbers and types of sensors. The evaluation result suggests a positive correlation between error and distance from Tobii: nearby, error is often less than 10mm and sometimes even sub-millimeter horizontally and vertically; at greater distances, error is often above 100mm.

Fig. 1 plots error against distance from the Tobii sensor, using data collected across several test sessions. The many outliers come from at least three sources: sensor noise, fast movement of the user's head, and interference of the FARO arm with Kinect facial recognition. We removed most of the data with FARO interference, but some of it remained (see the blue dashed ovals in Fig. 1). The black dashed box highlights positions near Kinect's minimum range, where its error grows rapidly. Kinect's tracking becomes more stable beyond 1200mm. Tobii can sometimes track past its designed range, but it produces samples much less frequently. In general, Tobii is more reliable at close range and is often the only source of tracking, whereas at a farther range Kinect becomes the better source.

Table 1 lists the root mean squared errors (RMSE) of Tobii and Kinect data vs. FARO over distance. Kinect gains accuracy past 800 millimeters, whereas Tobii's accuracy degrades past 1050 millimeters. For the fused tracking system, we divided the tracked range into three parts: under 800mm, we report Tobii positions; from 800 to 1050mm, we average Tobii and Kinect positions; and for farther distances, we report Kinect positions. The fused result has the lowest error in the overlapping range as is shown in Fig. 1 and Table 1.

5 Conclusion & Future Work

We have constructed a system that automatically fuses the commodity Kinect and Tobii tracker outputs, producing a new sensor with a combined range large enough to support emerging display usage scenarios, such as low-latency remote conferencing at a standing desk, with both user and display moving relative to one another.

Although we find that Kinect has better accuracy at a larger distances, we have not tested all of its operating range, because the FARO arm could not span it. Our results suggest that our fused tracker will work well beyond the range we tested. Because the FARO arm interfered with Kinect tracking at times, we might improve our evaluation with a different "gold standard" tracker. While our sensor fusion clearly improves the tracked range, we need further work to improve accuracy when the Tobii and Kinect ranges overlap, and neither is at its best: Kinect is disturbed by fast head movement, while Tobii samples intermittently. Rather than simply averaging, we might use a distance-based interpolation, regression, or filter to improve results, sidestepping the sharp modal behavior of our three-part Tobii-only, overlapping, and Kinect-only scheme.

Acknowledgments

We thank the University of North Carolina at Chapel Hill for providing the FARO arm, and Andrei State and Jim Mahaney for their assistance in setting up the device. As always, Turner Whitted provided valuable inspiration and advice. This work was supported in part by a grant from the National Science Foundation.

References

[1] J. Kim, G. Park, Y. Kim, S.-W. Min, and B. Lee. Elimination of image discontinuity in integral floating display by using adaptive image mapping. *Applied optics*, 48(34):H176–H185, 2009.

[2] O. Kinect. Open kinect.

[3] Microsoft Corporation. *Kinect for Windows Human Interface Guidelines v2.0*, 2014.

[4] R. Rohling, P. Munger, J. M. Hollerbach, and T. Peters. Comparison of relative accuracy between a mechanical and an optical position tracker for image-guided neurosurgery. *Journal of image guided surgery*, 1(1):30–34, 1995.

[5] T. Svoboda, D. Martinec, and T. Pajdla. A convenient multicamera self-calibration for virtual environments. *Presence: Teleoperators & virtual environments*, 14(4):407–422, 2005.

[6] Tobii Technology. *Positioning in front of an eye tracker*, 2016.