A Systematic Evaluation of Multi-Sensor Array Configurations for SLAM Tracking with Agile Movements

Brian M. Williamson*
University of Central Florida

Eugene M. Taranta II†
University of Central Florida

Pat Garrity‡
Army Research Laboratory

Robert Sottilare§
Army Research Laboratory

Joseph J. LaViola Jr. ¶
University of Central Florida

Abstract

Accurate tracking of a user in a marker-less environment can be difficult, even more so when agile head or hand movements are expected. When relying on feature detection as part of a SLAM algorithm the issue arises that a large rotational delta causes previously tracked features to become lost. One approach to overcome this problem is with multiple sensors increasing the horizontal field of view. In this paper, we perform a systematic evaluation of tracking accuracy by recording several agile movements and providing different camera configurations to evaluate against. We begin with four sensors in a square configuration and test the resulting output from a chosen SLAM algorithm. We then systematically remove a camera from the feed covering all permutations to determine the level of accuracy and tracking loss. We cover some of the lessons learned in this preliminary experiment and how it may guide researchers in tracking extremely agile movements.

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities ; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Tracking

1 Introduction

Accurate environment mapping and localization are critical components of an augmented reality (AR) system [1]. Simultaneous localization and mapping (SLAM) algorithms that track agents through an unknown, marker-less environment, while simultaneously constructing a map of the same area, offers one possible solution. One challenging application of SLAM in this domain involves tracking a human agent over an agile movement, using only affordable RGB-D sensors or a mobile device. Agile movements are characterized as rapid, high frequency motion, such as those encountered when one spins around, waves a flag, or ducks behind cover [4].

Since localization methods often match image-based feature descriptors across frames, high descriptor quantities enable better pose estimation. Although commonly employed descriptors are invariant to scale, position, and lighting conditions, as Castle et al. [2] note, agile human movement results in significant tracking error due to motion blur compounded with a loss of features after extreme rotation deltas. As a result of these issues, most SLAM systems suffer high error rates during agile movement [3], or they simply lose tracking. One option to resolve this is by combining collocated camera streams into a single information source, as described by Williamson et al. for AgileSLAM [4]. By increasing coverage via the use of multiple camera sources, localization procedures are better able to tolerate losses during agile movement.

Although straightforward in concept, there are many open questions. How many cameras are required to maintain tracking throughout an agile movement? Will two cameras back-to-back work as well as four cameras arranged in a square? What is the trade-off between tracking accuracy and camera count? Is there a relationship between movement type and camera configuration? To address these questions and others, we have started a systematic evaluation of multiple RGB-D sensor configurations over varying agile movements and present preliminary results that show what accuracy levels may be possible with fewer sensors than an array that provides a 360-degree horizontal field of view.

2 Experiment and Results

To begin our evaluation we recorded data similar to [4]. New data was recorded rather than using the existing data set so that IMU information could be included for future experiments and more agile movements could be added to the data set as part of our final evaluation. The code originally created to process this data was incorporated for our evaluation and is detailed in [4].

We recorded data from an array of ZED sensors, which capture RGB-D information via two stereo cameras, that were held by a user via a wooden post that extended beneath the sensors. We recorded seven calibration movements to verify that truth data had the same frame of reference as the SLAM estimate. These movements were slow and methodical along each axis (X, Y and Z) and rotating along each axis (yaw/heading, pitch/attitude, roll/bank). The seventh calibration data set involved a slow turn 90 degrees to the right and a movement forward along the relative Z axis.

The agile data recorded contained the following motions. ”Fast-180” was a 180 degree turn and an equally fast turn back. ”Turn and Duck” involved a simultaneous 90 degree turn while ducking. ”FPS Simulator” contained large translation movements that may be seen in a physically active video game. ”Eye Track” had large rotation deltas of someone following a fast virtual object with their eyes or controller. ”Flag Controller” had the sensor array represent a flag and undergoing motions that may be seen as such.

The SLAM algorithm we tested against reconstructs a feature’s world point using equation 1 from [4]. We modified this code to base the reconstruction depending on the permutation of the camera configuration we were running. For example, one configuration does not incorporate the backward facing camera, so the rotation of the features would only occur in front of the user, to the right of the user and to the left of the user. This allowed our modified images to still be correctly processed by the algorithm.

Each frame recorded was provided to the SLAM algorithm which would output an estimated pose (position and orientation) and statistics about the frame. The system could also advise us if an estimate should be rejected which we did so and added to a rejected frame counter. Since the four camera configuration was considered the ideal scenario we looked at the percentage of frames it considered good for each data set and marked an X for any configuration

*e-mail: brian.m.williamson@knights.ucf.edu
†e-mail: etaranta@gmail.com
‡e-mail: patrick.j.garrity4.civ@mail.mil
§e-mail: robert.a.sottilare.civ@mail.mil
¶e-mail: jjl@cs.ucf.edu

Table 1: RMSE Position (in centimeters) for each camera configuration by data set

Scenario	4	3RO	3LO	3FO	3BO	2RL	2RB	2FR	2FL	2FB	2BL	1R	1L	1F	1B
X	5.27	6.53	6.83	5.64	5.08	7.77	12.43	7.85	X	X	X	63.80	X	X	X
Y	4.50	5.54	4.53	5.04	4.82	5.48	6.08	4.99	X	5.01	X	14.71	X	X	X
Z	3.97	6.29	4.10	4.47	3.18	10.68	5.27	4.06	9.32	6.48	X	182.38	X	8.43	X
Yaw	7.77	14.39	7.57	6.36	15.09	X	135.33	6.64	27.14	X	14.53	X	X	X	X
Pitch	13.38	9.76	12.08	15.63	15.41	17.53	14.26	12.24	X	6.92	X	26.41	X	X	X
Roll	17.55	18.25	14.33	17.12	22.52	X	10.96	12.57	27.24	15.08	X	X	X	X	X
Look & Z	8.69	13.50	10.03	13.54	14.97	X	X	18.36	22.84	X	21.50	X	X	X	X
Fast 180	5.17	21.54	30.39	26.76	24.98	23.85	X	X	X	30.49	X	X	X	X	X
Turn & D	10.09	17.22	15.12	X	12.83	X	X	X	X	X	X	X	X	X	X
FPS Sim	13.64	X	10.64	X	14.54	X	X	12.94	X	X	X	X	X	X	X
Eye Track	12.19	15.83	7.31	14.71	14.48	X	X	10.74	35.18	X	X	X	X	X	X
Flag	13.82	X	7.42	X	13.53	X	X	X	X	X	X	X	X	X	X

Table 2: RMSE Orientation (in degrees) for each camera configuration by data set

Scenario	4	3RO	3LO	3FO	3BO	2RL	2RB	2FR	2FL	2FB	2BL	1R	1L	1F	1B
X	3.74	3.32	3.70	3.79	3.72	3.88	3.49	3.66	X	X	X	8.95	X	X	X
Y	1.98	2.00	2.01	2.10	1.86	1.92	2.28	1.99	X	2.05	X	6.04	X	X	X
Z	2.31	2.31	2.28	2.44	2.35	3.04	2.28	2.45	2.29	2.28	X	136.82	X	2.51	X
Yaw	4.25	4.38	4.29	4.84	4.13	X	20.88	5.19	5.47	X	5.33	X	X	X	X
Pitch	9.59	8.25	9.82	9.77	9.46	9.55	9.92	7.11	X	6.91	X	24.94	X	X	X
Roll	9.24	9.23	9.51	8.36	8.73	X	9.21	6.90	7.91	9.69	X	X	X	X	X
Look & Z	6.41	6.17	6.90	7.48	6.51	X	X	7.33	5.61	X	4.30	X	X	X	X
Fast 180	9.15	11.62	10.60	12.87	12.42	11.71	X	X	X	10.98	X	X	X	X	X
Turn & D	2.69	2.84	3.64	X	2.33	X	X	X	X	X	X	X	X	X	X
FPS Sim	2.47	X	2.44	X	2.33	X	X	2.54	X	X	X	X	X	X	X
Eye Track	5.05	4.08	4.48	4.99	4.42	X	X	4.24	13.92	X	X	X	X	X	X
Flag	3.80	X	2.76	X	3.80	X	X	X	X	X	X	X	X	X	X

that failed to meet 75% of its performance.

In Table 1 and Table 2 we present the RMSE values from the pose estimate to truth data delta for each configuration by each data set. The configurations are abbreviated using the following symbols. F for the forward facing camera, R for the right facing camera, B for the back facing camera, L for the left facing camera, O for the camera that is switched off and the numbers 1-4 for the total number of cameras on. As such the configuration where three cameras are on and the back camera is off is labeled 3BO; where as two cameras on, right and left, are labeled 2RL.

3 Discussion

In this initial analysis we notice several interesting features that are worthy of discussion and further exploration.

For the number of cameras we found that the worse single performing cameras were the left (1L) and rear (1B) facing cameras. The left feed featured a large featureless dome unique to our lab environment while the rear camera was partially obscured by the user, an issue that would be expected of an AR system using the forward and rear cameras on a phone. Thus, practitioners will need to consider how a device is held relative to the user, and if effective use of a rear facing camera will require the device to be held in an unusual, nonstandard way.

With the two camera configurations we found that pairing worse feeds with better ones did not significantly reduce the results. Furthermore we found that which pair is ideal (perpendicular or 180 degrees apart) depends on the motion being performed. We see our largest improvements when, from the start of an action to its end, the view of a first camera falls into the view of a second. In regards to three cameras we saw the performance was almost identical to the four camera configuration.

Furthermore, camera placement relative to expected motion is important. Consider pitch, which is rotation about the right facing axis. In this scenario, the front camera view changes dramatically and this contributes to why, when compared to a Z-axis translation, tracking is worse. In our evaluation, all cameras were aligned on the same plane so that when held upright, the forward direction of each camera sat parallel to the floor. Based on this relationship between rotation and camera orientation, it may be beneficial to adjust certain cameras so that their forward directions are perturbed up or down as we expect that most agile motions will involve a yaw component.

4 Conclusion and Future Work

This work represents our initial venture into multiple camera configurations and the effect that it has on localization accuracy. In future work we would like to incorporate more sensor configurations into our recording beyond the square formation. These may include two cameras with some overlap and cameras that are facing slightly upwards or downwards. We also intend to run through several other SLAM algorithms to help determine more generalized results.

While this line of questioning may be in its preliminary stages, we feel it will be essential for the development of practical augmented reality applications with head mounted displays or hand held devices.

Acknowledgements

This work is supported in part by NSF Award IIS-1638060, Lockheed Martin, Office of Naval Research Award ONRBAA15001, and Army RDECOM Award W911QX13C0052.

References

[1] R. T. Azuma. A survey of augmented reality. *Presence: Teleoperators and virtual environments*, 6(4):355–385, 1997.

[2] R. Castle, G. Klein, and D. W. Murray. Video-rate localization in multiple maps for wearable augmented reality. In *2008 12th IEEE International Symposium on Wearable Computers*, pages 15–22, Sept 2008.

[3] J. J. LaViola Jr, B. M. Williamson, R. Sottilare, and P. Garrity. Analyzing slam algorithm performance for tracking in augmented reality systems. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, 2017.

[4] B. M. Williamson, A. Vargas, P. Garrity, R. Sottilare, and J. J. LaViola Jr. Agileslam: A localization approach for agile head movements in augmented reality. In *Adjunct Proceedings of the International Symposium on Mixed and Augmented Reality*, pages 25–30, 2018.