Creating a 360-Degree RGB-D Sensor System for Augmented Reality Research

Brian M Williamson, Joseph J. LaViola Jr.
University of Central Florida
Orlando, Florida
Brian.M.Williamson@knights.ucf.edu,
JJL@eecs.ucf.edu

Robert Sottilare, Pat Garrity
U.S. Army Research Laboratory-Human Research and Engineering Directorate
Orlando, FL
robert.a.sottilare.civ@mail.mil,
patrick.j.garrity4.civ@mail.mil

ABSTRACT

Augmented Reality systems require both localization of the user and mapping of the surrounding area in order to correctly display virtual objects in a manner that is believable to the user. A single sensor can accomplish this with a SLAM algorithm but faces issues if the user performs a significant and quick rotation of the head as features that were being tracked are lost. An omnidirectional camera (360-degree horizontal, near 180-degree vertical) can resolve this, but COTS solutions in this domain only provide RGB information. In this paper we demonstrate a prototype system that fuses the imagery of four RGB-D sensors to create a 360-degree horizontal sensor feed of both color and depth information. We detail the design of the sensor array and challenges faced when attempting to record or visualize the data in real time with each camera's frame synchronized to other information necessary for future experiments. We also discuss the fusion system used and how this can detect features as a user rotates the sensor array in motions similar to human head movements. In the end, our sensor array shows the potential for quick, COTS-based prototype units that may make use of two or more RGB-D sensors in order to provide accurate localization and mapping of the environment for augmented reality research.

ABOUT THE AUTHORS

Joseph J. LaViola Jr. is the Charles N. Millican Faculty Fellow and Associate professor in the Department of Electrical Engineering and Computer Science and directs the Interactive Systems and User Experience Research Cluster of Excellence at the University of Central Florida. He is the director of the modeling and simulation graduate program and is also an adjunct associate research professor in the Computer Science Department at Brown University. His primary research interests include pen-based interactive computing, 3D spatial interfaces for video games, human-robot interaction, multimodal interaction in virtual environments, and user interface evaluation. His work has appeared in journals such as ACM TOCHI, IEEE PAMI, Presence, and IEEE Computer Graphics & Applications, and he has presented research at conferences including ACM CHI, ACM IUI, IEEE Virtual Reality, and ACM SIGGRAPH. He has also co-authored "3D User Interfaces: Theory and Practice," the first comprehensive book on 3D user interfaces. In 2009, he won an NSF Career Award to conduct research on mathematical sketching. Joseph received a Sc.M. in Computer Science in 2000, a Sc.M. in Applied Mathematics in 2001, and a Ph.D. in Computer Science in 2005 from Brown University.

Dr. Robert A. Sottilare leads adaptive instructional science programs at the US Army Research Laboratory where the focus of his research is automated authoring, instructional management, and analysis tools and methods for intelligent tutoring systems (ITS) design. His work (over 170 technical papers) is widely published and includes journal articles in the AI in Education, Cognitive Technology, Educational Technology & Society, and the Journal for Defense Modeling & Simulation. He is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT), an open source, AI-based adaptive instructional architecture. He is the lead editor for the Design Recommendations for Intelligent Tutoring Systems book series and the founding chair of the GIFT Users Symposia. He is a program committee member, and frequent speaker at the Defense & Homeland Security Simulation, Augmented Cognition, and AI in Education conferences. He is a member of the AI in Education Society, the Florida AI Research Society, and the IEEE Standards Association. He is a faculty scholar and adjunct professor at the University of Central Florida where he teaches a graduate level course in ITS design. Dr. Sottilare is also a frequent lecturer at the United States Military Academy (USMA) where he teaches a senior level colloquium on adaptive training and ITS design. He has a long history of participation in international scientific fora including NATO and the Technical Cooperation Program. Dr. Sottilare is the recipient of the Army Achievement Medal for Civilian Service (2008), and two lifetime achievement awards in Modeling & Simulation: US Army RDECOM (2012) and National Training & Simulation Association (2015).

Brian M. Williamson is a faculty researcher at the University of Central Florida in the department of Electrical Engineering and Computer Science within the Interactive Systems and User Experience Lab under Dr. LaViola. His primary research includes 3D user interfaces with previous papers written on the RealNav system, a natural locomotion system evaluated for video game interfaces. At the University of Central Florida, Brian has published a thesis on natural locomotion, and has been published by IEEE and CHI.

Pat Garrity is a Chief Engineer at the U.S. Army Research Laboratory-Human Research and Engineering Directorate, Advanced Simulation Technology Division (ARL-HRED-ATSD), He currently works in Dismounted Soldier Simulation Technologies conducting research and development in the area of dismounted soldier training and simulation where he is the Army's Science and Technology Manager for the Augmented Reality for Training Science and Technology Objective (STO). His current interests include Human-In-The-Loop (HITL) networked simulators, virtual and augmented reality, and immersive dismounted training applications. He earned his B.S. in Computer Engineering from the University of South Florida in 1985 and his M.S. in Simulation Systems from the University of Central Florida in 1994.

Creating a 360-Degree RGB-D Sensor System for Augmented Reality Research

Brian M Williamson, Joseph J. LaViola Jr.,
University of Central Florida
Orlando, Florida
Brian.M.Williamson@knights.ucf.edu,
JJL@eecs.ucf.edu

Robert Sottilare, Pat Garrity
U.S. Army Research Laboratory-Human Research
and Engineering Directorate,
Orlando, FL
robert.a.sottilare.civ@mail.mil,
patrick.j.garrity4.civ@mail.mil

INTRODUCTION

Augmented reality has been a growing research domain, particularly regarding training of psychomotor skills (Hughes, Stapleton, Hughes, & Smith, 2005). However, for the virtual elements to have a believable location in the real world, accurate tracking and environment mapping is necessary. One solution is a marker-based environment which has shown commercial success (HTC, 2018), but requires a pre-configured environment and is limited to the detection of the markers and their location. A marker-less environment is more versatile but can struggle to localize the user during extreme agile head movements (LaViola, Williamson, Sottilare, & Garrity, 2017). Our research goal is to improve the marker-less solutions to support the degree of movements that may be seen in dismounted soldier training.

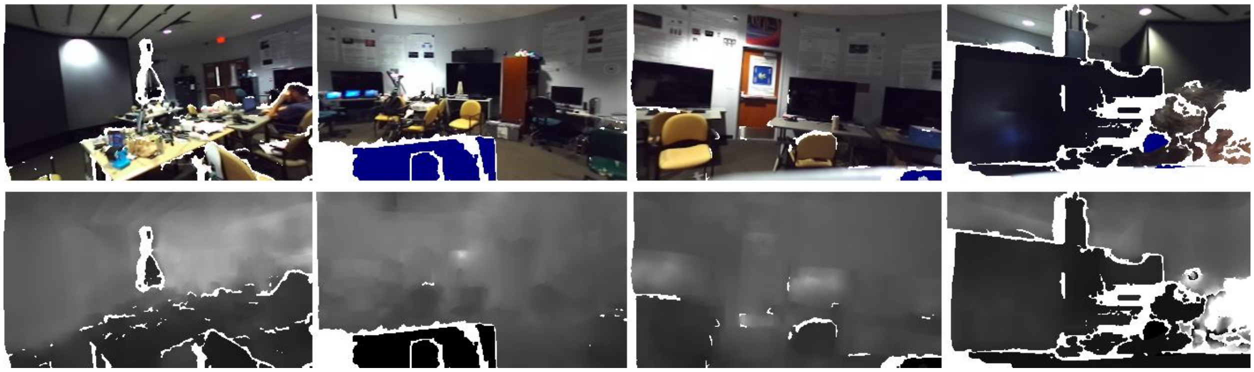

Figure 1. Example of fused color and depth imagery from RGB-D sensor array.

A common solution to marker-less environments in augmented reality research is the Simultaneous Localization and Mapping (SLAM) algorithm (Azuma, 1997). This algorithm comes from the robotics field and uses computer vision (Davison & Murray, 2002) to both detect the user's location and map out the environment. Some algorithms focus on color information only (Engel, Schöps, & Cremers, 2014) while others make use of color and depth sensors (Endres, et al., 2012). As can be seen, the sensor used for environment detection becomes an integral part of the research.

In this paper, we propose the construction of a sensor array that can provide accurate color and depth data in a 360-degree horizontal field of view. This will allow ample information to be available to the computer vision portion of SLAM algorithms even while undergoing extreme agile head movements. We also go through the framework developed to process the information from the sensor array and how the setup can be used for algorithm development and evaluation. Figure 1 shows an example of the fusion of color (RGB) and depth information pulled from the RGB-D sensors used in our array.

In the next section we discuss related research. In section three we go through our sensor array's development and justification. Section four demonstrates how the array can be used for algorithm development and evaluation. Section five concludes the paper and discusses future work.

RELATED WORK

The use of virtual or mixed (augmented) reality for the training of psychomotor tasks, especially for dismounted training has a long history. In Witmer, Bailey and Knerr (1995), dismounted soldier training in a virtual environment was examined, not just for effectiveness, but also covering concepts such as immersion and motion sickness. This research was expanded in Knerr, Lampton, Thomas, Corner and Grosse (2003), which conducted an extensive experiment regarding mission rehearsal training in virtual environments. It was determined in the experiment that training could take place in a virtual environment and it was predicted that as technology improved, so would the training capabilities of such environments. In Knerr and Lampton (2005), virtual environments were used for training in military operation in urban training (MOUT) and was determined to be effective in training with an advantage over live training in the variety of environments that could be presented with reduced preparation times.

Regarding mixed reality there is also a history of research for dismounted solider training. In Livingston, et al., (2002), the Battlefield Augmented Reality System (BARS) was developed to demonstrate augmented reality in MOUT training scenarios. In Hughes, Stapleton, Hughes and Smith (2005), several examples of mixed reality were evaluated, including MOUT training that made use of blue screen technology to place virtual avatars. Hughes notes that augmented reality can have an advantage over a purely virtual environment due to haptic and visual feedback provided by real objects. Recent developments have shown useful applications for augmented reality beyond just training, such as Winer and Schlueter (2017), which looks at using augmented reality for in the field expert assistance for equipment repairs using augmented reality. In Cisneros, Castillo, Johnson, Baker, & Garrity (2017) one of the many technical problems in augmented reality research are analyzed, in this case the issue of dynamic occlusion, which is to ensure the virtual system becomes aware of a new object, such as a person, stepping into the environment.

Simultaneous Localization and Mapping (SLAM) algorithms began as a solution in the robotics domain as a means for a robot to both understand its environment and its own location within that environment using simple sensors, such as a laser range finder (Azuma, 1997). This was expanded with visual SLAM, which would make use of cameras and computer vision algorithms (Davison & Murray, 2002). Since then the research has continued to grow, with several algorithms meant to solve specific applications. For example, the parallel tracking and mapping (PTAM) (Klein & Murray, 2007) algorithm was designed to use a single RGB sensor to provide accurate tracking of a desktop workspace. This would later be expanded to a system that utilized the Oriented FAST and Rotation BRIEF (ORB) (Rublee, Rabaud, Konolige, & Bradski, 2011) algorithm of feature detection and performed room wide detection, known as ORB-SLAM (Mur-Artal, Montiel, & Tardos, 2015). These systems would generally rely on stereo estimations between frames to determine depth information of the environment, but that can prove inaccurate if the number of common features between frames are low.

RGB-D SLAM (Endres, et al., 2012) incorporated a sensor that provides both color and depth information for each pixel to create a robust system that creates an accurate color point cloud along with the tracking of the user. ORB-SLAM would also expand to incorporate an RGB-D sensor for more accurate environment mapping (Mur-Artal & Tardos, 2016). While the addition of depth information improves environment mapping, it is still dependent on accurate localization, which when undergoing extreme agile movements, especially rotation changes, can become unreliable (LaViola, Williamson, Sottilare, & Garrity, 2017).

SENSOR ARRAY DEVELOPMENT

In our previous research (LaViola, et al., 2015; LaViola, Williamson, Sottilare, & Garrity, 2017) we realized that accurate environment mapping is dependent on accurate localization, and that can be problematic in a marker-less environment when large rotation deltas take place. Every SLAM algorithm depends upon its capability to compare the current frame of data with some previous frame, usually an established keyframe. Similar features (dense or sparse) are identified between the two frames and the changes are analyzed to estimate the change of the user's position. However, if few similar features exist between frames, an estimate cannot be determined. In this regard, rotation deltas are the most problematic as features can become lost within milliseconds if the user turns their head quickly. We designed our sensor array to solve this problem by ensuring features are always present no matter how much the user rotates.

There are other solutions to consider for the problem as well. A sensor could have an improved framerate to capture deltas in rotation fast enough, but this would have to be a very high framerate as the human head can move with burst speeds up to 780 degrees per second (Grossman, Leigh, Abel, Lanska, & Thurston, 1988). The images would also need to be free of blur and shearing to accomplish this and the processing of the frames would need to approach the speed of the camera. These difficulties may be overcome, but at a potentially high cost. Another solution is to rely on an inertial measurement unit (IMU) that can provide the missing rotation data, but the translation data that it can estimate tends to not be reliable as the accelerometer data can be noisy during fast movements. Furthermore, the IMU solution would not be able to track a rotation and translation movement, such as a user turning quickly and ducking at the same time.

Considering this, we proceeded with our solution to construct a sensor array, which would be at a lower cost than a high-speed camera and more reliable than a single sensor with an IMU. The array would be constructed of four ZED cameras (Stereo Labs, 2018) arranged in a square formation. The ZED sensor, shown in Figure 2, provides RGB-D information, but unlike other sensors which use a scanning laser and detect the range of the reflection, the ZED uses the stereo estimate of two cameras set a fixed distance apart. This produces depth information with an accuracy range of 0.5 meters to 20 meters (Stereo Labs, 2018), which is ideal for room scanning which would be necessary in our augmented reality applications. Furthermore, it had variable resolution settings and at the lowest setting (672 by 376) the field of view is nearly 90 degrees horizontal.

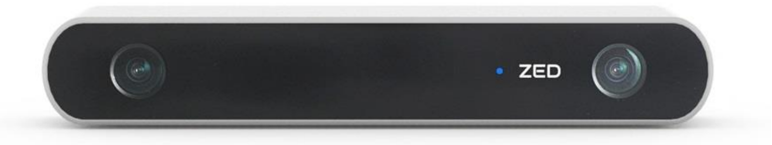

Figure 2. ZED sensor used for sensor array. The two cameras are at a calibrated fixed distance to produce accurate stereo depth data.

We used a 3-D printer to create platforms for each camera to rest on which could then be mounted to a more advanced frame if desired. Figure 3 shows the arrangement of the sensors with the fourth sensor removed to show the 3-D printed platform used. We considered this a prototype unit that while large for the average human head, it provided us with a proof of concept that could be improved upon by removing unnecessary components and casings from each sensor. At the time of the sensor array's construction the ZED Mini was not yet available, but we intend for future iterations to make use of that sensor.

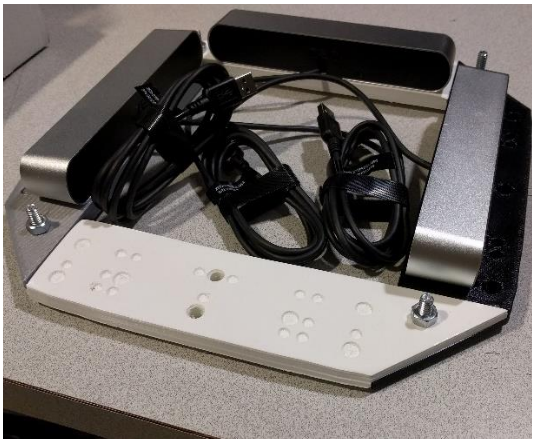

Figure 3. Arrangement of ZED sensors on 3-D printed platform. Fourth sensor is removed to show an example of the platforms.

Each sensor requires 380mA of power carried through the USB line, as such a powered USB hub was used to connect each sensor to make sure adequate current was available. We also had to consider the bandwidth needs, even with USB 3.0's capability of 3.2 Gbps (400 MB/s). The ZED hardware returns two frames, side by side, which contain both color and depth information encoded into 16 bits. In Table 1 we present the resolution options for each ZED camera and the required bandwidth needed for a single sensor and for our sensor array at either the maximum framerate or 30 frames per second. Given our single powered USB hub we tested with the lower resolution (672 by 376) at 30 frames per second. By utilizing more USB hubs, or installing a PCI-E USB 3.0 expansion card, higher resolution/framerate combinations would have been possible.

Table 1. Bandwidth Requirements for ZED Sensor Array

Resolution	Framerate	Single Camera Bitrate	Sensor Array Bitrate
2K	15 (max)	1.316 Gbps	5.265 Gbps
1080	30	1.99 Gbps	7.962 Gbps
720	30	884 Mbps	3.539 Gbps
VGA	30	242 Mbps	970 Mbps

EXPERIMENTAL DESIGN

With the sensor array designed we proceeded with a data recording experiment that would save off information to be used for comparison of SLAM algorithms similar to other data sets developed in the research domain (Geiger, Lenz, Stiller, & Urtasun, 2013) (Nardi, et al., 2015). We decided for the experiment to also include a Kinect sensor mounted above one of the ZED cameras to represent a traditional single sensor approach. The rationale for this was that several SLAM algorithms had been evaluated against the Kinect, thus its data could be used in comparison to the sensor array. We also incorporated an HTC Vive Tracker (HTC, 2018) to the center of our system to provide accurate truth data. The HTC Vive makes use of a marker-based tracking system, which has shown itself to have high accuracy so long as the light houses are visible to the tracking system.

Due to the weight of the increased sensors we decided the system would not be head mounted, but rather hand-held and head movements would be simulated. To accomplish this, we mounted the sensors to a wooden platform and attached a dowel through the center, as shown in Figure 4.

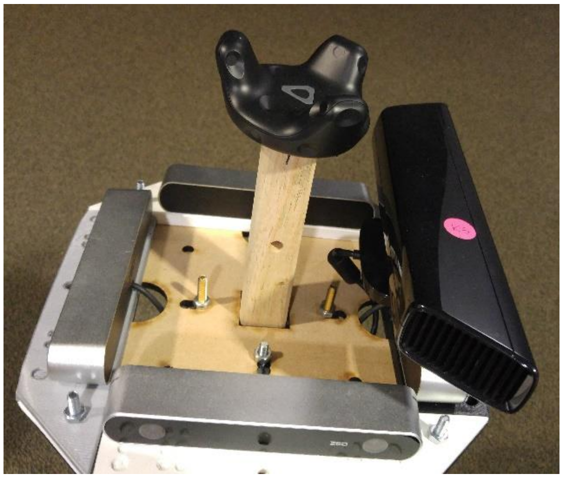

Figure 4. Mounting of sensor array with Kinect and Vive tracker for data recording sessions.

For processing of the data, we ran into a conflict with the choice of using multiple ZED sensors and a Kinect sensor on the same platform. The Kinect and VIVE software development kits (SDKs) that we had access to required Windows, while the ZED SDK's multi-sensor support was only available on Linux. We resolved this by making use of two laptops networked together, one which ran the Windows operating system while the other ran Ubuntu. The laptops used a simple TCP connection where the Ubuntu laptop functioned as the data recording and the

Windows laptop functioned as a streamer. Once the Ubuntu laptop established a connection, it would receive Kinect frames and truth data that it could then record.

We next had to consider the issue of synchronizing all the data in a threaded system. There was a thread created for each of the four ZED cameras and one thread created for the TCP connection to the Windows laptop for Kinect and truth data. If allowed to run asynchronously, each thread would have a variable number of frames grabbed and recorded which would then have to be synchronized by examining time windows.

Instead we opted for a state machine approach where the main thread of the recording program would dictate to the other five threads the state of the system. This would transition between a frame grab state, where each thread would grab its next frame of data either from the sensor or the TCP connection, followed by a record state where each frame would write to the disk and notify the main thread when it was done. Once every thread was complete, the system would be allowed to proceed to the next frame. Figure 5 shows this state machine.

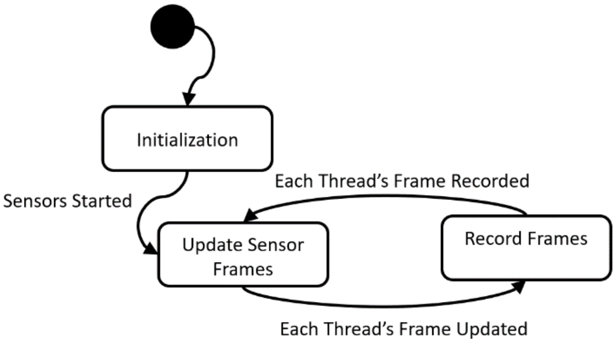

Figure 5. Thread State Machine used for synchronized frame recording.

We found the system recorded at roughly ten frames per second being largely limited by the speed of the traditional hard drive. A solid-state drive would have likely given us improved performance and will be considered for all future iterations. Picture information was not altered nor compressed before being written to the hard drive, making an individual "frame" of data from a recording session roughly 18MB in size.

We then went through a series of recording sessions that mimicked the movements of the human head, which could provide raw data files to be processed later by each SLAM algorithm. As a proof of concept, however, we found the sensor array to be successful in producing high quality frames where the same features could be observed between frames with large user movements, regardless of the speed at which the user moved. We did notice blurring during movement transitions and attributed this primarily to the extreme movements and the use of low quality resolution frames. Figure 6 shows how a movement transition can result in extremely blurry frames while moving, but features can still be recognized once the movement ends.

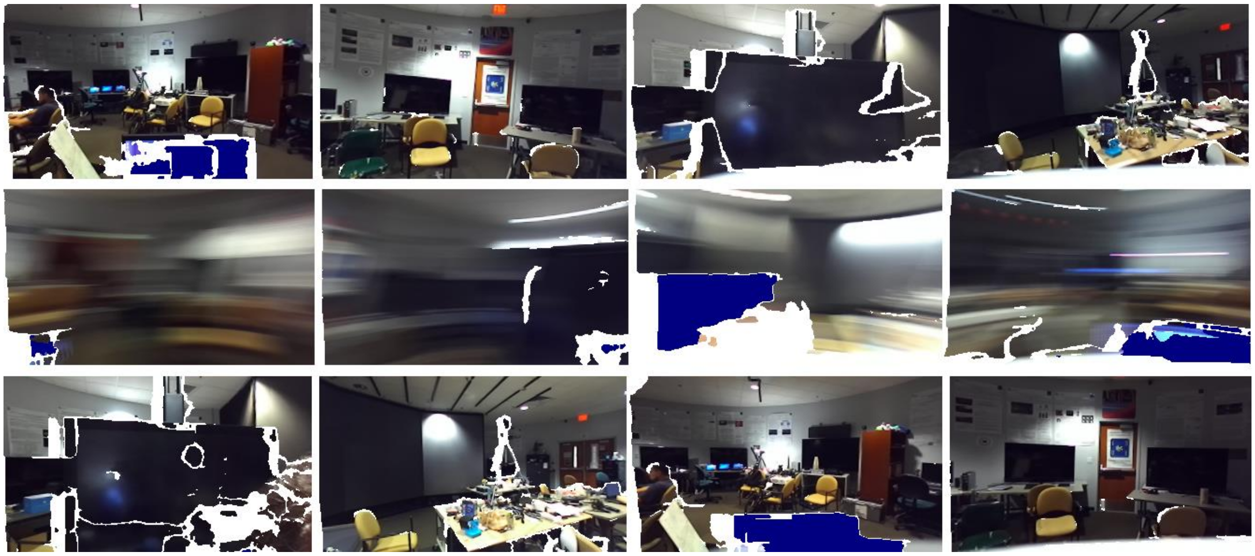

Figure 6. Example of extreme agile movements. The top frame is a normal color frame, the middle is during a 180-degree transition and the bottom set is once the transition has ended.

CONCLUSION AND FUTURE WORK

In this paper we demonstrated the construction of a prototype sensor array to be used for SLAM algorithm comparison. We showed both the creation of the sensor array and justification for its use in augmented reality research. Furthermore, this paper shows an example application of recording from such a sensor array and issues that may arise, such as USB bandwidth and frame synchronization. We also demonstrate issues that can arise with such a sensor array design and recommendations for overcoming them, primarily the bandwidth needs driving the resolution we were able to use.

In our future work we intend to run the data recorded through multiple SLAM algorithms, including one designed to incorporate the 360-degree horizontal field of view generated by the sensor array. We considered the use of two omni-directional cameras to find stereo data where they overlap but felt issues may still arise if the user turns quickly 90 degrees and begins to look at areas where depth data could not be determined. Still, we intend to attempt a direct comparison of a stereo omni-directional camera approach and our sensor array.

While our focus was on augmented reality and correcting visual registration of virtual objects, it is possible that this improved tracking and registration could also reduce simulator sickness as it would correct visual-kinesthetic and visual-proprioceptive errors (Azuma, 1997). In future work, we would like to examine the improvements our system brings to virtual environments regarding simulator sickness through an extensive user study.

We would also consider making use of the new ZED Mini in future prototypes. At the current specs, our system measures 7in by 8in with a weight of 1.5lbs and with the ZED Mini this would be reduced to 5in by 6in with a weight of 0.6lbs.

Furthermore, our sensor array is only 360-degrees in the horizontal field of view, and features may become lost if the user looks up quickly. Multiple sensors could alleviate this, but it also represented an unnatural movement to increase the pitch angle of the head to such a degree quickly, as opposed to yaw changes when a user turns around.

Finally, our sensor array did not incorporate an IMU and we did not consider an IMU sufficient by itself to solve this problem. However, while transitioning between scene and experiencing a large amount of blur, an IMU combined with a system such as a Kalman filter may be able to provide accurate estimates even when features are temporarily unavailable. Once the system settles to a location, corrections could then be applied as familiar features are determined.

Our sensor array is a prototype to be considered for marker-less tracking in augmented reality research. While it has large requirements in terms of bandwidth and processing speed, it provides accurate RGB-D data with a nearly 360-degree horizontal field of view which shows promising in feature-based tracking even after extreme agile head rotations.

REFERENCES

Azuma, R. (1997). A survey of augmented reality. *Presence: Teleoperators and virtual environments*, *6*(4), 355-385.

Cisneros, J., Castillo, J., Johnson, S., Baker, J., & Garrity, P. (2017). Simulation – Dynamic Occlusion using Fixed Infrastructure for Augmented Reality. *Interservice/Industry Training, Simulation, and Education (I/ITSEC).* Orlando.

Davison, A. J., & Murray, D. W. (2002). Simultaneous localization and map-building using active vision. *IEEE transactions on pattern analysis and machine intelligence*, *24*(7), 865-880.

Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., & Burgard, W. (2012). An evaluation of the RGB-D SLAM system. *Robotics and Automation (ICRA)*, 1691-1696.

Engel, J., Schöps, T., & Cremers, D. (2014). LSD-SLAM: Large-scale direct monocular SLAM. *European Conference on Computer Vision*, 834-849.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 1231-1237.

Grossman, G. E., Leigh, J. R., Abel, L., Lanska, D. J., & Thurston, S. (1988). Frequency and velocity of rotational head perturbations during locomotion. *Experimental brain research*, *70*(3), 470-476.

HTC. (2018, May 25). *Vive Pro*. Retrieved from Vive web site: https://www.vive.com/us/product/vive-pro/

Hughes, C., Stapleton, C., Hughes, D., & Smith, E. (2005). Mixed reality in education, entertainment, and training. *IEEE computer graphics and applications, 25*(6), 24-30.

Klein, G., & Murray, D. (2007). Parallel tracking and mapping for small AR workspaces. *ISMAR*, 225-234.

Knerr, B. W., & Lampton, D. R. (2005). *An assessment of the virtual-integrated MOUT training system (V-IMTS).* Orlando: ARMY RESEARCH INST FOR THE BEHAVIORAL AND SOCIAL SCIENCES.

Knerr, B. W., Lampton, D. R., Thomas, M., Corner, B. D., & Grosse, J. R. (2003). *Virtual environments for dismounted soldier simulation, training, and mission rehearsal: Results of the FY 2002 culminating event.* . Orlando: ARMY RESEARCH INST FIELD UNIT.

LaViola, J. J., Williamson, B., Brooks, C., Veazanchin, S., Sottilare, R., & Garrity, P. (2015). Using augmented reality to tutor military tasks in the wild. *Interservice/Industry Training, Simulation, and Education (I/ITSEC)*, 1-10.

LaViola, J. J., Williamson, B., Sottilare, R., & Garrity, P. (2017). Analyzing SLAM Algorithm Performance for Tracking in Augmented Reality Systems. *Interservice/Industry Training, Simulation, and Education (I/ITSEC).* Orlando.

Livingston, M. A., Rosenblum, L. J., Julier, S. J., Brown, D., Baillot, Y., Swan, J. E., . . . Hix, D. (2002). *An augmented reality system for military operations in urban terrain.* Washington D.C.: NAVAL RESEARCH LAB.

Mur-Artal, R., & Tardos, J. D. (2016). ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *arXiv*.

Mur-Artal, R., Montiel, J. M., & Tardos, J. D. (2015). ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics, 31*(5), 1147-1163.

Nardi, L., Bodin, B., Zia, M., Mawer, J., Nisbet, A., Kelly, P. H., . . . Furber, S. (2015). Introducing SLAMBench, a performance and accuracy benchmarking methodology for SLAM. *Robotics and Automation (ICRA)*, 5783-5790.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *2011 International conference on computer vision*, (pp. 2564-271).

Stereo Labs. (2018). *Meet the Zed Stereo Camera*. Retrieved from Stereo Labs: https://www.stereolabs.com/zed/

Winer, E., & Schlueter, J. (2017). Expert-Assisted Field Maintenance using Augmented Reality. *Interservice/Industry Training, Simulation, and Education (I/ITSEC).* Orlando.

Witmer, B., Bailey, J., & Knerr, B. W. (1995). *Training Dismounted Soldiers in Virtual Environments: Route Learning and Transfer.* Orlando: ARMY RESEARCH INST FOR THE BEHAVIORAL AND SOCIAL SCIENCES.