Simultaneous Personnel Localization and Mapping

A Major Qualifying Project Report Submitted to the Faculty of
Worcester Polytechnic Institute

In partial fulfillment of the requirements for the
Degree of Bachelor of Science in Electrical & Computer Engineering

By:

Georges Gauthier
John DeCusati

May 6, 2016

Advisor:
Professor R. James Duckworth

Contents

List of Figures

Abstract

The overall goal of this project is to create a device capable of generating detailed maps and imagery of an area in real-time. This device will rely on an FPGA, and will use image processing algorithms capable of detecting, localizing, and tracking human beings. The device will gather imagery from the visual light and infrared-spectrums, as well as localization data and distance measurements from an IMU and a rangefinder. Along with gathering and processing data, the device will also serve as a long-range wireless access point, and will be able to transmit all generated maps and imagery in real-time. A major deliverable of this project is that the transmitted data will be fully processed, allowing it to be viewed remotely on less-powerful, mobile devices.

This device will be especially useful for first responders. It is intended to be mounted on a small remote control vehicle, allowing any connected user to wirelessly traverse dangerous and remote locations in search of people in need. Since this device transmits data in real-time, it will be able to provide first responders with an accurate representation of not only a 2-D floor plan of an area such as a building, but also where any people are located. An anticipated use of this device would be in the event of a building in danger of collapsing. Since it would be dangerous to physically enter the building, first responders could locate any people trapped inside and find the fastest route to them using the wirelessly transmitted floorplan. The first responders would also be aware of any dangers in their way by making use of the real-time augmented video stream. This video stream will consist of image data with overlaid with object indicators and location information on any human beings detected by the image processing algorithms.

1 Introduction

Currently there are many applications that rely on a simple video camera setup in order to gather information on remote and inaccessible locations. Although this is an effective strategy for simple surveillance, it is limited in many ways. Using current imaging and sensor technology, it is possible to gather camera images and 3D depth information on a given area as both a cost-effective and information-rich alternative to using a camera module on a device. If a product were to be created that gathers this information as a replacement to using a standalone camera module, it would be possible to use high speed data processing techniques in both hardware and firmware that would allow for the creation of an augmented real-time video feed.

This type of technology is known as Simultaneous Localization And Mapping, or SLAM. The purpose of SLAM is to compute the location of an agent within its environment, and allows for the creation of self-aware robot systems that are able to respond to their surroundings. SLAM is a common area of research in the image processing and high-speed computing field, and has been applied mainly to autonomous vehicles. We would like to propose the creation of a SLAM-like system that is capable of monitoring and mapping its environment in real-time, as well as detecting and localizing objects, such as human beings. For the purposes of our project, we would like to define our desired objective as Simultaneous Personnel Localization And Mapping, or SPLAM.

A device that is capable of both mapping its surroundings using SLAM and performing human detection in real-time would be applicable to many different fields. We are especially interested in creating a proof of concept sensor suite capable of performing these tasks that can be added to existing robotic systems as a stand-in replacement for a video camera. This type of technology would allow for people such as firefighters or first responders to wirelessly traverse dangerous and remote locations in search of people in need. We envision our sensor suite being able to process data so that its users would be provided with a 2D ?floorplan? of the area being traversed by the sensor suite, as well as an augmented video feed with

imagery containing indicators for any detected human beings in the area.

One type of technology that would be useful for performing the high speed data processing necessary for SPLAM is a Field Programmable Gate Array , or FPGA. FPGAs pose several advantages over using standard computing or microcontroller technology for real-time data processing, as they have the ability to manipulate digital information in parallel using hardware only. This allows for extremely high-speed performance, as calculations can be run in parallel and are only dependent on their data inputs as opposed to waiting for specific tasks or scripts to run on a microcontroller or computer software interface.

Although FPGA technology is highly applicable to performing SLAM-like tasks due to its high speed, there are currently few existing commercial products that use FPGAs for the purpose of performing SLAM. Most current SLAM implementations rely on the use of a sensor suite connected to a computer or system on chip (SoC) computing device that performs data analysis using software or a real-time operating system (RTOS). This means that data must first be collected by a sensor suite, and then transferred to an external computing device that is only capable of processing it serially based on its arrival time. Although this type of setup is acceptable for performing real-time situational awareness analysis, we propose that an embedded, FPGA-based SPLAM device would be a much more elegant and higher-speed solution.

2 Background

A major concern with real-time image processing, especially in first responder situations, is speed. Because FPGAs have the ability to process data in parallel, they are ideal for this type of application. Using an FPGA for this system will enable all data inputs to be processed at the same time, thereby dramatically increasing throughput speed. Dealing with each input separately makes it easier to combine everything together, especially because each component functions at a different clock speed. Also, since everything is running in parallel, more cameras can be added to the system to increase the field of vision of the device without introducing any latency in the system, as long as enough memory is available.

SLAM is a widely expanding field with much potential for improvement. One application of such a system is a proof of concept of camera-based SLAM systems, presented by Andrew Davison of Oxford University in a research paper entitled "Real-Time SLAM with a Single Camera" [1]. This system is handheld and relies on a computer using a 2.2 GHz Pentium processor connected to a single camera and laser rangefinder. The system requires prior knowledge of the area being analyzed before it can successfully localize and map. It implements edge detection, but is limited to the narrow field of vision of the rangefinder, so it is only able to map an object directly in front of it. This system carries a latency of around 33 milliseconds. An output frame of the device is shown in Figure 1.

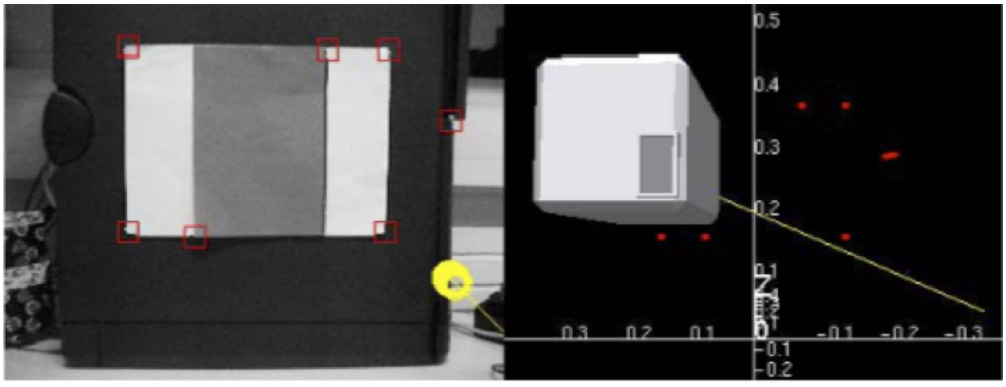

Figure 1: Real-Time SLAM with a Single Camera [1]

The frame on the left in Figure 1 is the video feed with 6 points of a paper target input as

prior knowledge, along with successfully marked identifying features (marked as red squares), and another identifying feature that is not marked for measurement (marked by a yellow circle). The frame on the right is a localization graph displaying the positions of all red squares.

A more commercial device similar to our concept is Serveball?s Squito^{TM} [2]. Squito is a wireless, throwable, 360° panoramic camera that implements target detection to stabilize the video feed from its many cameras. It is shown in Figure 2 below.

Figure 2: Serveball's Squito [2]

Squito utilizes a microprocessor receiving input from cameras, as well as orientation and position sensors in order to transmit a real-time stabilized video of its adventure. The device is still in the prototype stage and is receiving interest from first responders. The image in Figure 3 shows the input from the Squito's four cameras on the left, and the corresponding stitched output on the right.

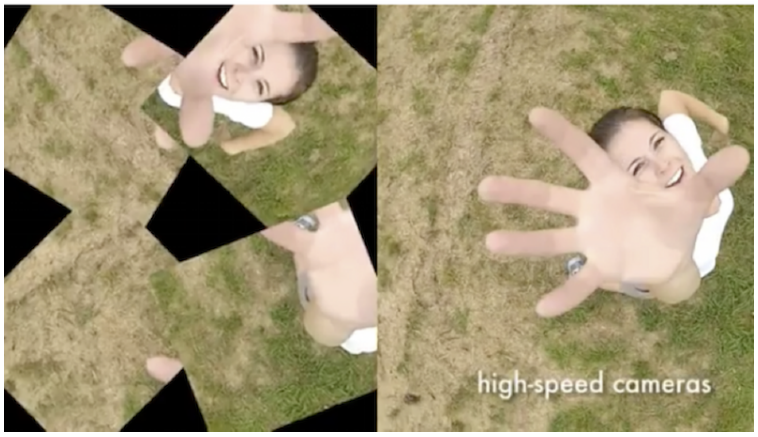

Figure 3: Serveball's Squito Input and Output [2]

By using multiple camera sensors in a sensor suite, it is also possible to determine depth information from corresponding images of an area. This technique is known as stereo imaging, and the process of gathering depth information from a pair of stereo images is known as disparity mapping. University of Bologna researchers Stefano Mattoccia and Matteo Poggi have worked to implement a real-time disparity mapping algorithm on an FPGA, and an example of a stereo image disparity is shown in Figure 4 below [3]. Using their stereo vision algorithm, the researchers are able to generate real-time image data showing the relative locations of objects within an image frame using color gradients. Based on this depth information, it is also possible to detect objects located within the field of view of the stereo imaging system, as shown in Figure 4. An implementation similar to this would be extremely useful in a SLAM-like system, as it would allow for the localization of objects and creation of 2D "floorplans" of an area in real-time using only two camera sensors.

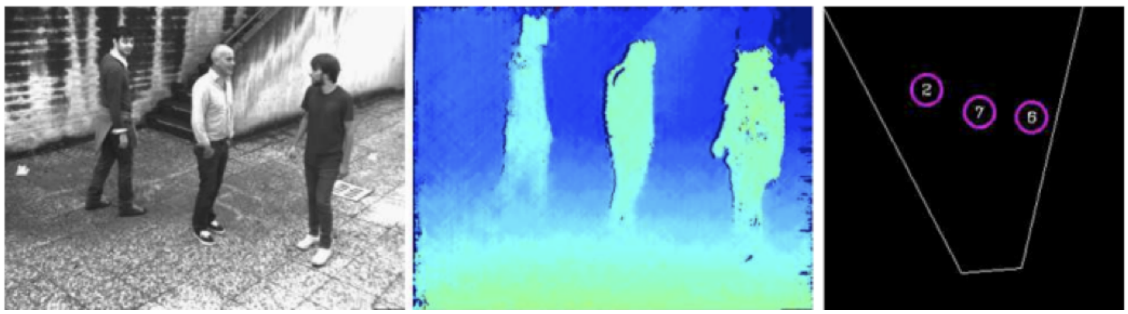

Figure 4: From Left to Right: Original Image, Disparity Map, Object Detection Results [3]

Many security systems implement human detection and human body tracking in order to increase their effectiveness. These devices process real-time images in order to identify human characteristics, and are limited to the field of vision of a stationary or rotating camera. An example of this type of system is explored by the Mitsubishi corporation in a research paper entitled "Human Body Tracking by Adaptive Background Models and Mean-Shift Analysis" [4]. The paper explores a stationary image processing system implemented on a PC platform with a 1.8GHz processor that yields a maximum processing time of 100 milliseconds. An output frame of the system is shown in Figure 5 below.

Figure 5: Human Body Tracking by Adaptive Background Models and Mean-Shift Analysis [4]

Our proposed device will combine the ideas of the four systems examined. It will be able to simultaneously localize and map an area, as well as implement human detection algorithms. The device will be capable of generating real-time 2D maps of any area it has traversed with humans' locations labeled, and an augmented video feed of what the device is recording with any humans' positions marked.

In order to successfully implement this system, we propose the creation of a device that will rely on two stereo cameras, a laser rangefinder, and an inertial measurement unit (IMU) as its sensor suite, as shown in Appendix Item 2. Limitations of previous art have been in their ability to combine human detection with real-time localization and mapping of a

large field of vision. Little to no existing commercial products are also capable of processing their gathered data locally and in real-time, with their gathered data usually requiring post-processing on external computing devices. Stereo cameras will allow our device to calculate disparity, just as human eyes do. Although disparity is useful for localization, it is not enough for accurate mapping because it only accurately provides the relative distance between objects. The inclusion of a rangefinder will allow for precise base distance readings, and an IMU will be used to spatially reference all gathered data. All of this data will be combined with the disparity maps and image data in order to create flawless localization and mapping. All time-dependent processing required for the device will be mainly done in parallel using hardware on an FPGA. An overall functional block diagram of our intended implementation is shown in Figure 6 below.

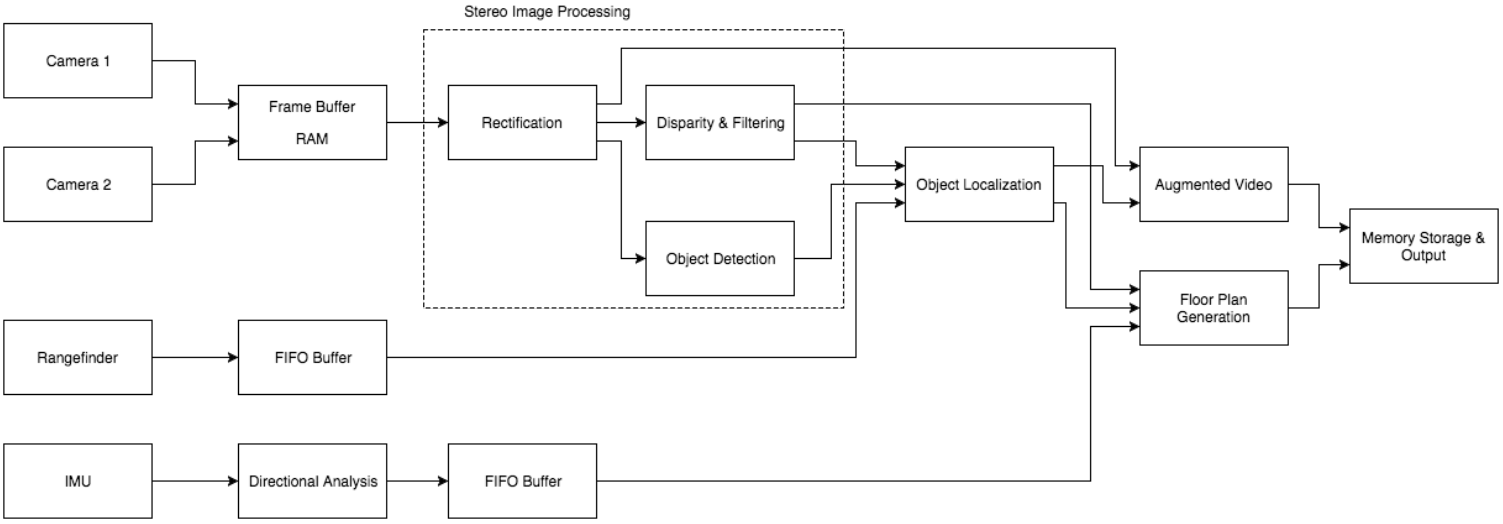

Figure 6: Functional Block Diagram

Most applicable previous camera-based systems have also focused on object detection from a stationary point, or edge detection from a mobile platform. Our project aims to combine these concepts, by creating a mobile device that detects people which will be especially of use in many first responder situations. In addition, this device will receive data from the visible light and infrared spectrums in order to identify people quickly and accurately in a way that has not been previously implemented.

As our research has progressed over time, our project objectives have continually evolved. We originally envisioned the creation of a device that used laser rangefinders to create 3D maps of its surroundings, similar to that of a Carnegie Mellon University device created in order to volumetrically map abandoned mines [5].

As our research progressed, we believed that we could use a visual light and thermal imaging camera set to gather information on an area, and supplement that data with IMU and rangefinder readings in order to produce detailed maps of our sensor suites? surroundings. Eventually we came upon the concept of disparity mapping and generating depth information from image data, and decided that we would again like to shift the overall setup of our device to rely mainly on stereo image data. Due to our overall budget and the resources that have been made available to us, in the coming terms we plan to use an electronic rangefinder, IMU, and stereo camera pair to generate real-time SPLAM video and floorplan information. Although we were also originally planning on including a thermal camera in our sensor suite as well, we have decided to eliminate the module in favor of higher quality cameras due to its prohibitive cost, low resolution, slow sampling rate, and small field of view. More information on this decision can be found in Appendix Item 3.

References

[1] Davison, A.J., *Real-Time Simultaneous Localisation and Mapping with a Single Camera*. IEEE Computer Vision, 2003. 2(1).

[2] *Serveball*. Available from: `http://www.serveball.com/`.

[3] Stefano Mattoccia, M.P., *A passive RGBD sensor for accurate and real-time depth sensing self-contained into an FPGA*. in *International Conference on Distributed Smart Cameras*. 2015.

[4] Fatih Porikli, O.T., *Human Body Tracking by Adaptive Background Models and Mean-Shift Analysis*. in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. 2003.

[5] Sebastian Thrun, D.H., David Ferguson, Michael Montemerlo, Rudolph Triebel, and C.B. Wolfram Burgard, Zachary Omohundro, Scott Thayer, William Whittaker. *A System for Volumetric Robotic Mapping of Abandoned Mines*. in *IEEE International Conference on Robotics and Automation*. 2003.

3 Appendix

3.1 Useful Resources

This section is intended to serve as complete compilation of all resources gathered throughout D Term 2016 that we believe will be useful as we begin to work on the methodology portion of our project.

- Strother, Daniel. 2011. "Open-Source FPGA Stereo Vision Core Released." https://danstrother.com/2011/06/10/fpga-stereo-vision-core-released/.
- Field, Mike. 2013. "Zedboard OV7670." http://hamsterworks.co.nz/mediawiki/index.php/Zedboard_OV7670.
- "OV7670/OV7671 CMOS VGA CameraChip Implementation Guide." https://www.fer.unizg.hr/_download/repository/OV7670new.pdf.
- "MIPI CSI2-to-CMOS Parallel Sensor Bridge - Lattice Semiconductor." http://www.latticesemi.com/~/media/LatticeSemi/Documents/ReferenceDesigns/JM/MIPICSI2toCMOSPar pdf?document_id=50533.
- "OmniVision Serial Camera Control Bus (SCCB) Functional Specification." http://www.ovt.com/download_document.php?type=document&DID=63".
- Morvan, Yannick. "Multiple-View Depth Estimation." http://www.epixea.com/research/multi-view-coding-thesisse15.html.
- MathWorks. "Depth Estimation From Stereo Video." http://www.mathworks.com/help/vision/examples/depth-estimation-from-stereo-video.html.
- Stefano Mattoccia, Matteo Poggi. 2015. "A passive RGBD sensor for accurate and real-time depth sensing self-contained into an FPGA". International Conference on Distributed Smart Cameras.

- Mattoccia, Stefano. 2013. "Stereo Vision: Algorithms and Applications." http://www.slideshare.net/DngNguyn43/stereo-vision-42147593.

- Szeliski, Richard. 2010. Computer Vision: Algorithms and Applications. New York: Springer.

- Beau Tippets, Dah Jye Lee, Kirt Lillywhite, James Archibald. 2013. "Review of Stereo Vision Algorithms and Their Suitability for Resource-Limited Systems." Journal of Real-Time Image Processing 11 (1). doi: 10.1007/s11554-012-0313-2.

- Jouni Rantakokko, Joakim Rydell, Peter Stromback, Peter Handel, Jonas Callmer, David Tornqvist, Fredrik Gustafsson, Magnus Jobs, Mathias Gruden. 2011. "Accurate and Reliable Soldier and First Responder Indoor Positioning: Multisensor Systems and Cooperative Localization." IEEE Wireless Communications 18 (2):10-18. doi: 10.1109/MWC.2011.5751291.

- Davison, Andrew J. 2003. "Real-Time Simultaneous Localisation and Mapping with a Single Camera." IEEE Computer Vision 2 (1). doi: 10.1109/ICCV.2003.1238654.

- Fatih Porikli, Oncel Tuzel. 2003. "Human Body Tracking by Adaptive Background Models and Mean-Shift Analysis." IEEE International Workshop on Performance Evaluation of Tracking and Surveillance.

- Giovanni Pintore, Enrico Gobbetti. "Effective mobile mapping of multi-room indoor structures." The Visual Computer 30 (6):707-716. doi: 10.1007/s00371-014-0947-0.

- "iRobot 110 FirstLook." iRobot. http://www.irobot.com/$~$/media/Files/Robots/Defense/FirstLook/iRobot-110-FirstLook-Specs.pdf.

3.2 Component Selection

Component	Part Number	Supplier	Cost
FPGA	ZedBoard	Borrowed	N/A
IMU	ADIS16375	Borrowed	N/A
Rangefinder	URG-04LX	Borrowed	N/A
Stereo Cameras†	MT9V034	Mouser	$146

† Note that we originally planned to purchase a flir lepton thermal camera module and accompanying breakout board to support two stereo ov7670 camera modules. After experimenting with the ov7670 camera module on our FPGA board, we began to realize that these camera modules are highly limited due to their low frame rate and poor documentation, and realized that we wanted to search for a different camera module. In addition, at a price of $223 for a thermal camera with an 80x60 degree resolution, 25 degree fov, and 7-9Hz image sample rate, we believe that we are much better off spending our money on better camera modules that will be more usable for our task. For more information see Appendix Item 3.

3.3 Camera Module Decision Matrix

Camera Module	Max Frame Rate (FPS)	Resolution at Max Frame Rate (px.)	Cost	Requires External Adapter	Data Transfer Interface	Shutter	Field of View (deg.)	Rank 0-10
OV7670	30	640x480	$10	No	Parallel	Rolling	25	5
Raspberry Pi Camera	90	640x480	$30	Yes, $53	MIPI (CSI2)	Rolling	49	6
PC1089K	60	720x480	$32	No	NSTC/ PAL	Rolling	Not Given	5
OV4682	330	640x480	$89	Yes, $50	MIPI	Rolling	Not Given	6
MT9V034	**60**	**750x480**	**$73**	**No**	**Parallel**	**Global**	**55**	**9**

For purposes of comparison, the thermal camera module we were evaluating is also shown below.

Flir Lepton	9	80x60	$175	Yes, $40	SPI/ MIPI	N/A	50	3

Shown above is our decision matrix for choosing a camera module. Fields marked in green indicate a positive ranking, while red indicates a negative ranking. Based on the individual

rankings of each item's field, we gave our camera modules an overall ranking of 0-10 in the right hand column, with 10 being an extremely high ranking and 0 being an extremely low ranking.

Based on our decision matrix, we believe that it would be worth both our time and money to use the MT9V034 camera modules for our stereo camera interface. These camera modules are the only low-cost global shutter option we've come across, and would be ideal for taking images in a sensor suite that is susceptible to motion. The MT9V034 also uses a parallel data interface and relies on an external clock and shutter trigger, making the module ideal for interfacing with an FPGA-based stereo imaging setup.

3.4 LaTeX Coding Examples

This section isn't intended to remain here, but can serve as an example for how to set things up later on

3.4.1 Figures

Figure 7: A Test Figure

Using the `\ref` command, I'm able to reference Figure 7 by calling `\ref{wpiLogo}`.

3.4.2 Code Snippet

Code snippets can be created by calling `\begin{lstlisting}`, inserting all code, and then calling `\end{lstlisting}`. Also call `\singlespacing` before the code snippet and `\doublespacing` after to keep things from getting too big.

```
1 //verilog code example
2 always @(x, y, z)
3   x <= y + z;
```

3.4.3 Using the bibliography

All bibliographic references are contained in `bib.tex`. To cite a reference in the paper, use the `\cite` command.

As an example, I can cite *Serveball* at the end of this sentence by calling `\cite{serveball}`.[2]

To cite multiple references, call `\cite{ref1,ref2}`.[2, 4]