Tyzx DeepSea High Speed Stereo Vision System

John Iselin Woodfill, Gaile Gordon, Ron Buck

Tyzx, Inc.
3885 Bohannon Drive
Menlo Park, CA 94025
{Woodfill, Gaile, Ron}@tyzx.com

Abstract

This paper describes the DeepSea Stereo Vision System which makes the use of high speed 3D images practical in many application domains. This system is based on the DeepSea processor, an ASIC, which computes absolute depth based on simultaneously captured left and right images with high frame rates, low latency, and low power. The chip is capable of running at 200 frames per second with 512x480 images, with only 13 scan lines latency between data input and first depth output. The DeepSea Stereo Vision System includes a stereo camera, onboard image rectification, and an interface to a general purpose processor over a PCI bus. We conclude by describing several applications implemented with the DeepSea system including person tracking, obstacle detection for autonomous navigation, and gesture recognition.

1. Introduction

Many image processing applications require or are greatly simplified by the availability of 3D data. This rich data source provides direct absolute measurements of the scene. Object segmentation is simplified because discontinuities in depth measurements generally coincide with object borders. Simple transforms of the 3D data can also provide alternative virtual viewpoints of the data, simplifying analysis for some applications.

Stereo depth computation, in particular, has many advantages over other 3D sensing methods. First, stereo is a passive sensing method. Active sensors, which rely on the projection of some signal into the scene, often pose high power requirements or safety issues under certain operating conditions. They are also detectable - an issue in security or defense applications. Second, stereo sensing provides a color or monochrome image which is exactly (inherently) registered to the depth image. This image is valuable in image analysis, either using traditional 2D methods, or novel methods that combine color and depth image data. Third, the operating range and Z resolution of stereo sensors are flexible because they are simple functions of lens field-of-view, lens separation, and image size. Almost any operating parameters are possible with an appropriate camera configuration, without requiring any changes to the underlying stereo computation engine. Fourth, stereo sensors have no moving parts, an advantage for reliability.

High frame rate and low latency are critical factors for many applications which must provide quick decisions based on events in the scene. Tracking moving objects from frame to frame is simpler at higher frame rates because relative motion is smaller, creating less tracking ambiguity. In autonomous navigation applications, vehicle speed is limited by the speed of sensors used to detect moving obstacles. A vehicle traveling at 60 mph covers 88 ft in a second. An effective navigation system must monitor the vehicle path for new obstacles many times during this 88 feet to avoid collisions. It is also critical to capture 3D descriptions of potential obstacles to evaluate their location and trajectory relative to the vehicle path and whether their size represents a threat to the vehicle. In safety applications such as airbag deployment, the 3D position of vehicle occupants must be understood to determine whether an airbag can be safely deployed - a decision that must be made within tens of milliseconds.

Computing depth from two images is a computationally intensive task. It involves finding, for every pixel in the left image, the corresponding pixel in the right image. Correct corresponding pixel is defined as the pixel representing the same physical point in the scene. The distance between two corresponding pixels in image coordinates is called the *disparity* and is inversely proportional to distance. In other words, the nearer a point is to the sensor, the more it will appear to shift between left and right views. In dense stereo depth computation, finding a pixel's corresponding pixel in the other image requires searching a range of pixels for a match. As image size, and therefore pixel density, increases, the number of pixel locations searched must increase to retain the same operating range. Therefore, for an NxN image, the stereo computation is approximately

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04 $ 20.00 IEEE

$O(N^3)$. Fortunately, the search at every pixel can be effectively parallized.

Tyzx has developed a patented architecture for stereo depth computation and implemented it in an ASIC called the DeepSea Processor. This chip enables the computation of 3D images with very high frame rates (up to 200 fps for 512x480 images) and low power requirements (< 1 watt), properties that are critical in many applications. We describe the DeepSea Processor in Section 2. The chip is the basis for a stereo vision system, which is described in Section 3. We then describe several applications that have been implemented based on this stereo vision system in Section 4 including person tracking, obstacle detection for autonomous navigation, and gesture recognition.

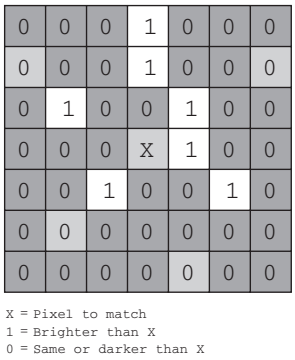

Figure 1: Census transform: pixels darker than center are 0's in bit vector, pixels brighter than center are 1's.

2. DeepSea Processor

The design of the DeepSea ASIC is based on a highly parallel, pipelined architecture [6, 7] that implements the Census stereo algorithm [8]. As the input pixels enter the chip, the Census transform is computed at each pixel based on the local neighborhood, resulting in a stream of Census bit vectors. At every pixel a summed Hamming distance is used to compare the Census vectors around the pixel to those at 52 locations in the other image. These comparisons are pipelined, with 52 comparisons occurring simultaneously. The best match (shortest summed Hamming distance) is located with five bits of subpixel precision. The DeepSea Processor converts the resulting pixel disparity to metric distance measurements using the stereo camera's calibration parameters and the depth units specified by the user.

Under specific imaging conditions, the search for correspondence can be restricted to a single scan line rather than a full 2D window. This simplification is possible, in the absence of lens distortion, when the imagers are coplanar, their optical axes are parallel, and corresponding scan lines are co-linear. The DeepSea processor requires rectified imagery (see Section 3) to satisfy these criteria with real-world cameras.

The DeepSea Processor also evaluates a number of "interest operators" and validity checks which are taken into account to determine the confidence of a measurement. One example is the left/right check. A correct measurement should have the same disparity whether the search is initiated in the left image or the right image. Different results indicate an invalid measurement. This check is expensive in software, but easily performed in the DeepSea Processor.

2.1. Advantages of the Census Transform

One problem that makes determining stereo correspondence difficult is that the left and right images come from distinct imagers and viewpoints, and hence corresponding regions in the two images may have differing absolute intensities resulting from distinct internal gains and biases, as well distinct viewing angles.

The DeepSea Processor uses the Census transform as its essential building block to compare two pixel neighborhoods. The Census transform represents a local image neighborhood in terms of its *relative intensity structure*. Figure 1 shows that pixels that are darker than the center are represented as 0's whereas pixels brighter than the center are represented by 1's. The bit vector is output in row major order. Comparisons between two such Census vectors are computed as their Hamming distance.

Because the Census transform is based on the relative intensity structure of each image it is invariant to gain and bias in the imagers. This makes the stereo matching robust enough that the left and right imagers can, for example, run independent exposure control without impacting the range quality. This is a key advantage for practical deployment.

Independent evaluations of stereo correlation methods [5, 2, 1] have found that Census performs better than classic correlation approaches such as normalized cross correlation (NCC), sum of absolute differences (SAD), and sum of squared differences (SSD).

2.2. DeepSea Processor performance

The figure of merit used to evaluate the speed of stereo vision systems is Pixel Disparities per Second (PDS). This is the total number of pixel to pixel comparisons made per second. This is computed from the area of the image, the width of the disparity search window in pixels, and the frame rate. The DeepSea Processor is capable of 2.6 billion PDS. This is faster than any other stereo vision system we are aware of by an order of magnitude or more. Additional performance details are summarized in Figure 2.

3. DeepSea Stereo System

The DeepSea Development System is a stereo vision system that employs the DeepSea Processor and is used to develop

DeepSea Processor Specifications	
Input Image Size (max)	512x2048 (10 bit)
Stereo Range	
Search Window	52 Disparities
Sub-pixel Localization	5 bits
Z output	16 bit
Max frame rate	200 fps (512x480)
Power	$<$ 1 watt

Figure 2: DeepSea Processor Specifications

Figure 3: DeepSea Board.

new stereo vision applications. The DeepSea Stereo Vision System consists of a PCI board which hosts the DeepSea Processor, a stereo camera, and a software API to allow the user to interact with the board in C++ from a host PC. Color and range images are transferred to the host PC's memory via DMA.

The DeepSea board (shown in Figure 3) performs communication with a stereo camera, onboard image rectification, and interfaces to a general purpose processor over the PCI bus. Since the host PC does not perform image rectification and stereo correlation, it is now available for running the user's application code.

The DeepSea stereo cameras are self-contained stereo cameras designed to work in conjunction with the DeepSea Board. DeepSea stereo cameras connect directly to the board using a high-speed stereo LVDS link. By directly connecting to the DeepSea system, latency is reduced and the host's PCI Bus and memory are not burdened with frame-rate, raw, image data. Tyzx has developed a family of stereo cameras which includes 5cm, 22cm, and 33cm lens separations (baselines) as shown in Figure 4. A variety of standard CMOS imagers are used based on application requirements for resolution, speed, color, and shutter type. Each camera is calibrated to define basic imager and lens parameters such as lens distortion and the exact relationship between the imagers. These calibration parameters are used by the system to rectify the images. After this transformation the images appear distortion free, with co-planar image planes, and with corresponding scan lines aligned.

Figure 4: Tyzx Stereo Camera Family: 5cm, 22cm, and 33cm baselines.

The frame rate of any given system configuration will vary based on the capabilities of the imagers. Common configurations include:

- Omnivision: image sizes 400x300 to 512x512, frame rates of 30fps to 60fps, color
- National: image sizes 320x240 to 512x480, frame rate of 30fps, high dynamic range
- Micron: image sizes 320x240 to 512x480, frame rates of 47fps to 90fps, Freeze frame shutter

4. Applications of Tyzx Stereo Sensors

Tracking people in the context of security systems is one application that is ideal for fast 3D sensors. The Tyzx distributed 3D person tracking system was the first application built based on the Tyzx stereo system. The fast frame rates simplify the matching of each person's location from frame to frame. The direct measurements of the 3D location of each person create more robust results than systems based on 2D images alone. We also use a novel background modeling technique that makes full use of color and depth data [3, 4] that contributes to robustness of tracking in the context of changing lighting. The fact that each stereo camera is already calibrated to produce absolute 3D measurements greatly simplifies the process of registering the cameras to each other and the world during installation. An example of tracking results is shown in Figure 5. The right image shows a plan view of the location of all the people in a large room; on the left these tracked locations are shown overlaid on the left view from each of four networked stereo cameras.

Fast 3D data is also critical for obstacle detection in autonomous navigation. Scanning laser based sensors have been the *de facto* standard in this role for some time. However, stereo sensors now present a real alternative because of faster frame rates, full image format, and their passive nature. Figure 6 shows a Tyzx DeepSea stereo system mounted on the Carnegie Mellon Red Team's Sandstorm

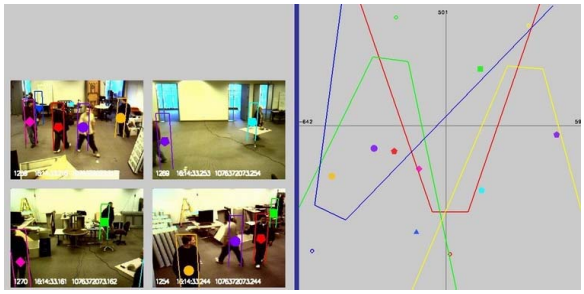

Figure 5: Tracking people in a large space based on a network of four Tyzx stereo integrated sensors.

Figure 6: CMU's Sandstorm autonomous vehicle. Tyzx DeepSea Stereo sensor is mounted in the white dome.

autonomous vehicle in the recent DARPA Grand Challenge Race.

Sensing and reacting to a user's gestures is a valuable control method for many applications such as interactive computer games and tele-operation of industrial or consumer products. Any interactive application is very sensitive to the time required to sense and understand the user's motions. If too much time passes between the motion of the user and the reaction of the system, the application will seem sluggish and unresponsive. Consider for example the control of a drawing program. The most common interface used for this task is a mouse - which typically reports its position 100 to 125 times per second. In Figure 7 we show an example in which a user controls a drawing application with her fingertip instead of a mouse. The 3D position of the figure tip is computed from a stereo range image and a greyscale image. The finger position is tracked relative to the plane of the table top. When the finger tip approaches the surface of the table, the virtual mouse button is depressed. Motion of the figure tip along the table is considered a drag. The finger moving away from the surface is interpreted as the release of the button. In this application, a narrow baseline stereo camera is used to achieve an operating range of 2 to 3 feet from the sensor with 3D spatial accuracy of ± 1.5 mm.

5. Future Directions and Conclusions

Tyzx high speed stereo systems make it practical to bring high speed 3D data into many applications. Integration

Figure 7: Using 3D tracking of the finger tip as a virtual mouse to control a drawing application. Inset shows left image view.

Figure 8: Tyzx Integrated Stereo System. Self contained stereo camera, processor, and general purpose CPU.

of the DeepSea Board with a general purpose processor creates a smart 3D sensing platform - reducing footprint, costs, power, and increasing deployability. Figure 8 shows a prototype stand-alone stereo system, incorporating a stereo camera, DeepSea Board, and a general purpose CPU. This device requires only power and ethernet connections for deployment. The processing is performed close to the image sensor, including both the computation of 3D image data and the application specific processing of these 3D images. Only the low bandwidth results, e.g. object location coordinates or dimensions, are sent over the network. Even higher performance levels and further reductions in footprint and power are planned. In the future we envision a powerful networked 3D sensing platform no larger than the stereo camera itself.

Acknowledgments

Tyzx thanks SAIC for providing the Tyzx stereo system to Carnegie Mellon for their Sandstorm Autonomous Vehicle.

References

[1] J. Banks, M. Bennamoun, and P. Corke. "Non-parametric techniques for fast and robust stereo matching," In *Proceedings of IEEE TENCON*, Brisbane, Australia, December 1997.

[2] S. Gautama, S. Lacroix, and M. Devy, "Evaluation of Stereo Matching Algorithms for Occupant Detection", in *Proceedings of the International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pages 177–184. Sept 1999. Cofus, Greece.

[3] G. Gordon, T. Darrell, M. Harville, J. Woodfill. "Background estimation and removal based on range and color", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (Fort Collins, CO), June 1999.

[4] M. Harville, G. Gordon, J. Woodfill, "Foreground Segmentation Using Adaptive Mixture Models in Color and Depth", *Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video*, (Vancouver, Canada), July 2001.

[5] Heiko Hirschmller, "Improvements in Real-Time Correlation-Based Stereo Vision", *Proceedings of IEEE Workshop on Stereo and Multi-Baseline Vision*, pp. 141-148. December 2001, Kauai, Hawaii

[6] J. Woodfill, B. Von Herzen, "Real-Time Stereo Vision on the PARTS Reconfigurable Computer," *Proceedings IEEE Symposium on Field-Programmable Custom Computing Machines*, Napa, pp. 242-250, April 1997.

[7] Woodfill, Baker, Von Herzen, Alkire, "Data processing system and method", U.S. Patent number 6,456,737.

[8] R. Zabih, J. Woodfill, "Non-parametric Local Transforms for Computing Visual Correspondence", *Third European Conference on Computer Vision*, (Stockholm, Sweden) May 1994.