SP1: Stereo Vision in Real Time

Konstantin Schauwecker[1]

Abstract— **Stereo vision is a compelling technology for depth perception. Unlike other methods for depth sensing, such as time-of-flight or structured light cameras, stereo vision is a passive approach. This makes this method suitable for environments with bright ambient lighting, or for situations with multiple sensors within close proximity to one another. The reason why stereo vision is not used more widely is that it requires a vast amount of computation. To overcome this burden, Nerian Vision Technologies introduces the SP1 stereo vision system. This stand-alone device is able to handle the required processing by relying on a built-in FPGA.**

I. INTRODUCTION

Stereo vision is one of the best researched areas in computer vision. Its origins date back to the 1970s, and it has since seen significant scientific advancement. Compared to other approaches for depth perception, stereo vision has the advantage of being a passive approach. Thus, unlike for structured-light or time-of-flight cameras, the performance of a stereo vision system is not affected by bright ambient lighting. Further, there is no interference if multiple stereo cameras observe the same space. Stereo vision is thus a compelling technology for systems that are expected to operate outdoors or within close proximity to one another.

Particularly for robotics, these properties are highly desirable. However, actual robots that make use of stereo vision are not very common. The key problem that has prevented a more wide spread adoption of this technology are the high computational demands. While many algorithms for stereo analysis have been proposed in literature within the past decades, good-performing methods are still very demanding in terms of processing resources, even on modern hardware.

While traditional CPUs struggle with the vast number of computations that are required for stereo vision, there exists another type of hardware that can handle this task much more efficiently, which are Field Programmable Gate Arrays (FPGAs). An FPGA is a generic integrated circuit that can be programmed to fulfill a particular application. Because programming is performed on the circuit level, an FPGA is not forced to follow the usual fetch-decode-execute cycle of a CPU. Rather, an application specific architecture can be found that divides the problem into many small sub-problems, which can each be solved in parallel.

This ability to massively parallel processing is what enables FPGA-based systems to gain high speed-ups when compared to equivalent CPU-implementations. At the same time, an FPGA requires significantly less power than a GPU when performing the same task, and FPGA-based systems can be built at a much smaller form factor.

Unfortunately, the effort involved in programming FPGAs vastly exceeds the effort of programming ordinary CPUs or GPUs. Thus, FPGAs are not yet commonplace in todays robotic systems. In order to make this technology more accessible to researches and product developers, Nerian Vision Technologies[2] has developed the SP1 stereo vision system, which is shown in Fig. 1a. Using an FPGA, this small-scale processing device is able to perform stereo vision in real-time and at high processing rates. Its key features are:

- Processes input images with resolution of 640 $\times$ 480 pixels from two connected USB industrial cameras.
- Covers a disparity range of 112 pixels, with a sub-pixel resolution of $1/16$ pixel.
- Can process up to 30 frames per second.
- Power consumption is below 4 W.
- Computed disparity map is transmitted through ethernet.

II. ARCHITECTURE

The stereo matching method that is implemented by the SP1 is based on Semi Global Matching (SGM) [1]. Since its proposal, SGM has enjoyed much popularity, due to its high quality results and its computational efficiency. SGM is at the heart of some of the best performing stereo matching methods, such as the one proposed by Žbontar et al. [2], which at the time of writing is the best performing stereo method on the KITTI vision benchmark suite.

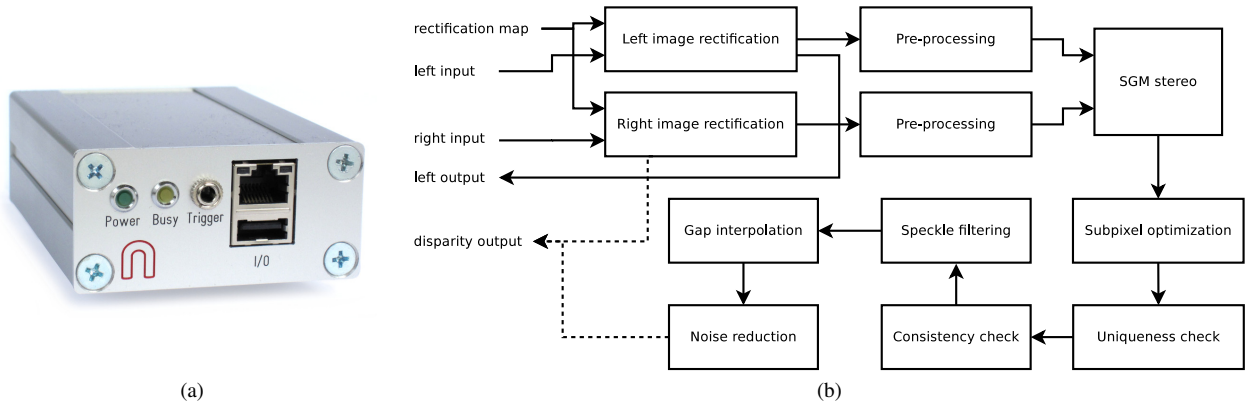
It has been shown that FPGA-based implementations of SGM are possible [3], [4]. However, only relying on SGM is not sufficient for receiving competitive stereo matching results. Rather, we require an entire image processing pipeline, which not only includes SGM, but also a range of different pre- and post-processing steps. We have implemented one such pipeline for the SP1, which is depicted in Fig. 1b.

In terms of pre-processing, the most important step is image rectification. This is a corrective image transformation that compensates for distortions from the cameras' optics and errors in the camera alignment. During rectification, the SP1 is able to move every image pixel by an offset of up to ± 31 pixel locations. Bi-linear interpolation is employed in order to support offsets with sub-pixel resolution.

After one further pre-processing step, which makes the subsequent methods more robust towards illumination variations, the SGM algorithm is applied. As a result, we receive a cost cube that assigns a matching cost to all possible

[1]K. Schauwecker is the founder of Nerian Vision Technologies, Gotenstr. 9, 70771 Leinfelden-Echterdingen, Germany.

[2]`http://nerian.com`



(a) (b)

Fig. 1: (a) SP1 stereo vision system and (b) block diagram of processing structure.

combination of left- and right-image pixel locations that are within the supported disparity range.

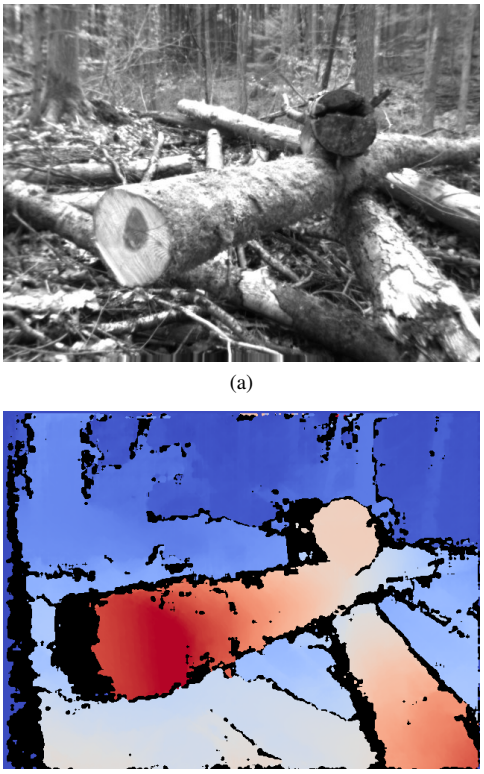
The cost cube is then passed through several post-processing steps. First, we try to improve the location accuracy of the best matching pixel pair through sub-pixel optimization. We then try to identify occlusions and mismatches through a uniqueness and a consistency check. The uniqueness check ensures that the matching cost for each selected pixel pair is significantly below the costs for other matching candidates. The consistency check enforces that the best matching pair is also selected if stereo matching is repeated in the opposite matching direction. Pixel locations that do not pass one of these checks are marked as invalid.

The remaining post-processing methods do not require the full cost cube. Hence, the cost cube can be reduced to a disparity map, which only encodes the horizontal offsets between pixel locations of corresponding left and right image pixels. We then identify small isolated speckles of similar disparity. Such speckles usually originate form false matches and we again mark the corresponding pixels in the disparity map as invalid. Small gaps of invalid pixels in the disparity map are then filled through interpolation. Finally, we apply a filter for noise reduction, as we expect the true depth of the scene to be somewhat smooth.

III. RESULTS AND CONCLUSION

An example for a left-camera input image and the corresponding disparity map, which has been computed by SP1, can be found in Fig. 2a and 2b. The disparity map can be easily converted into a depth map or a 3D point cloud. The SP1 delivers a stream of disparity maps through ethernet, at a rate of 30 frames per second. The latency for computing and delivering a disparity map is below the time interval between two frames (i.e. below 33 ms).

With the SP1 we have created a stand-alone system for stereo vision, which can easily be integrated into existing robots or other intelligent systems. We hope that by introducing the SP1, we will make stereo vision more accessible to a wide range of researchers and developers. This will hopefully facilitate the development of more systems that make use of the principles of stereo vision.



(a)

(b)

Fig. 2: Example for (a) left input image and (b) computed disparity map.

REFERENCES

[1] H. Hirschmüller, "Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information," in *IEEE Conf. on Comput. Vis. and Pattern Recogn. (CVPR)*, vol. 2, 2005, pp. 807–814.

[2] J. Žbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *IEEE Conf. on Comput. Vis. and Pattern Recogn. (CVPR)*, 2015.

[3] S. K. Gehrig, F. Eberli, and T. Meyer, "A real-time low-power stereo vision engine using semi-global matching," *Comput. Vis. Syst.*, pp. 134–143, 2009.

[4] C. Banz, S. Hesselbarth, H. Flatt, H. Blume, and P. Pirsch, "Real-time stereo vision system using semi-global matching disparity estimation: Architecture and FPGA-implementation," in *IEEE Int. Conf. on Embedded Comput. Syst. (SAMOS)*, 2010, pp. 93–101.