OFFSET APERTURE BASED HARDWARE ARCHITECTURE FOR REAL-TIME DEPTH EXTRACTION

Woojin Yun[1], *Young-Gyu Kim*[1], *Yeongmin Lee*[1], *Jinyeon Lim*[1], *Wonseok Choi*[1], *Muhammad Umar Karim Khan*[1], *Asim Khan*[1], *Said Homidov*[3], *Pervaiz Kareem*[1], *Hyun Sang Park*[2], *and Chong-Min Kyung*[1]

[1] School of Electrical Engineering, KAIST, Daejeon, Republic of Korea
[2] Division of Electrical Engineering, Kongju National University, Cheonan, Republic of Korea
[3] Center for Integrated Smart Sensors, Daejeon, Republic of Korea

ABSTRACT

Due to the increasing demand for 3D applications, development of novel depth-sensing cameras is being actively pursued. However, most of these cameras still face the challenge of high energy consumption and slow speed in the depth extraction process. This becomes a serious bottleneck in embedded implementations where real-time performance is required, constrained by power and area. This work proposes Offset Aperture (OA) camera, a new hardware architecture for fast, low-energy, and low-complexity depth extraction. Optimal implementations of pre-processing, cost-volume generation and cost-aggregation are presented. The whole depth-extraction pipeline has been implemented on a Field Programmable Gate Array (FPGA). Overall, a mere 2.8% of bad classification was achieved with the proposed system. Also, the proposed system can process 37 VGA frames per second while consuming 0.224 μJ/pixel. High accuracy, speed and low energy consumption of the proposed OA architecture make it suitable for embedded applications.

Index Terms— Depth extraction, FPGA, offset aperture, real time, low power.

1. INTRODUCTION

Depth extraction using stereo cameras has been pursued for many decades. However, the computational requirements for stereo and the lack of computational resources in the near past never allowed stereo vision to be used for commercial applications. Recently, stereo cameras have penetrated the market, thanks to the available computational capability of processors. This has led to a renewed interest in the research community towards novel depth extraction cameras generally, and in particular to accurate stereo cameras.

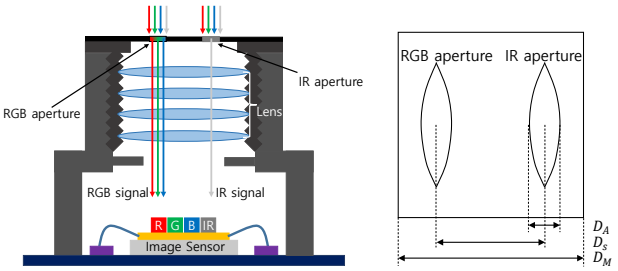
Although the computational capacity of modern digital hardware has significantly increased compared to the past, depth extraction using stereo cameras has been predominantly performed on PCs. Despite a huge demand, depth sensing cameras in mobile applications are yet to find its feet. This can be attributed to two main reasons. First, stereo cameras are too bulky for mobile applications. Second, the processing requirements of stereo vision are too high for mobile applications under real-time constraints.

Recently, the Offset Aperture (OA) camera was proposed [1] for depth sensing with a small footprint. It uses a single four-color sensor and two apertures, each with a spectrally different response. The two apertures are spatially displaced, resulting in the disparity in the observed image. Using modified apertures for depth extraction is not a new concept. The first such approach was proposed in [2] where the red and the green apertures are displaced, generating a depth dependent disparity across the red and the green channels. Similarly, displaced green, red and blue apertures have been proposed to estimate depth from an image in [3], [4] and [5]. However, unlike these approaches, OA provides an RGB image that is well-aligned spatially.

Although the OA camera pursues low computational power and small area on its own, it may lose its purpose if the depth extraction consumes significant computational power or area. In the past, numerous hardware implementations for stereo vision have been proposed such as [6], [7],[8], [9] and [10]. These implementations, however, cannot be used with OA as in OA the cost-volume is generated over spectrally different images, requiring a cost metric that is robust against scale and offset differences across the images.

This paper presents a new hardware architecture for depth extraction, applicable to cross-spectral stereo matching. Furthermore, the computational requirement of the proposed hardware is quite low. The proposed hardware architecture in conjunction with the OA camera will provide an ideal solution for depth extraction in mobile applications.

The rest of the paper is structured as follows. A brief review of the OA technology is presented in Section 2. Section 3 describes the hardware architecture. Experimental results are discussed in Section 4.



(a) (b)

Fig. 1. A graphical illustration of the OA camera system where (a) shows a structure of camera module and (b) shows the offset aperture. In this work, D_A = 1mm, D_S = 3.5mm and D_M = 6mm are used.

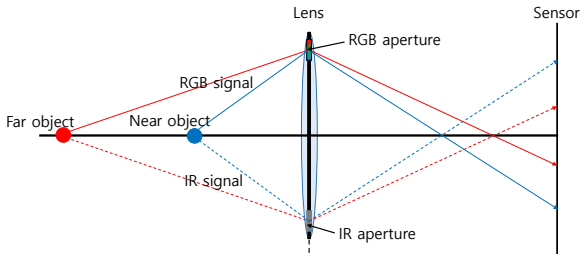


Fig. 2. The relationship between disparity and depth.

2. REVIEW OF OA CAMERA

Typically, the Bayer color filter array (CFA) is used in image sensors. A 2×2 grid on the Bayer CFA is composed of two green, one blue and one red pixel. IR energy to the sensor is blocked by a IR-blocking filter. On the contrary, the CFA used in this work replaces one green pixel on the 2×2 grid in the Bayer CFA by an IR pixel, and the IR-blocking filter is removed as shown in Fig. 1(a).

A unique aperture with distinct spectral characteristics is used in the OA camera. The aperture for RGB and IR images is horizontally displaced, resulting in horizontally offset RGB and IR images. This disparity is used to determine the depth of the scene. Graphical representation of the disparity-depth relationship is shown in Fig. 2. Note that cat's eye-shaped apertures in Fig. 1(b) are used for enhanced light efficiency while maximizing the baseline.

3. HARDWARE ARCHITECTURE OF THE DEPTH MAP PROCESSOR

A depth-map processor (DMP) for OA camera and other cross-spectral stereo applications is described in this section. Fig. 3 shows the top-level hardware architecture of the DMP. The DMP can be classified into two parts: the depth map engine (DME) which performs the core task of disparity estimation and the peripherals which provide necessary external interface. The DME extracts the depth (or disparity) from the stereo images which is displayed through the data transceiver. Furthermore, serial interfaces I2C and SPI are included to change the configuration of the DME based on the specific application. The tasks performed by DME are divided into three parts: pre-processing, main processing and post-processing. Note that the DME is fully pipelined to achieve high frame rate.

3.1. Pre-processing

Firstly, a raw image is interpolated to construct a full RGB-IR image in the image interpolation block. Since spectral responses of the color channels are not completely isolated from each other, a suppression of spectral cross-talk precedes other functional blocks. Under the assumption that dominant cross-talks in RGB channels are caused by IR band rays coming through IR aperture, the spectral cross-talk can be suppressed by subtracting weighted intensities of other channels. The weights represent degree of spectral interferences and, therefore, depend on the lighting condition. In order to estimate lighting condition, Bayesian approach in normalized RGB-IR color space is used [11]. After the cross-talk suppression, the gamma correction amplifies low RGB-IR signal values to secure them against bit truncation in the following blocks. For the suppression of the image noise, the mean filter is applied in the image noise reduction block.

3.2. Main processing

Generally, images are transformed before performing stereo matching. This enables better matching as certain features of the images are enhanced. In this work, we apply a cascade of transformations to the images before performing stereo matching.

Textureless areas have always been problematic for stereo matching. Here we perform gradient extraction on the images before matching, i.e., the texture is enhanced by extracting the one-dimensional first-order derivative of the images. Mathematically, the gradient images are obtained as

$$I'(x,y) = I(x+1,y) - I(x-1,y), \tag{1}$$

where $I(x,y)$ is the pixel intensity at the position (x,y) and $I'(x,y)$ is the gradient at the position (x,y). The gradient images are obtained for both the stereo images.

The second stage of transformations applies a local normalization to the stereo images. The images are normalized w. r. t. a local neighborhood Ω of $N \times N$ pixels as

$$I_N(x,y) = \frac{I'(x,y) - \mu_\Omega(x,y)}{\sigma_\Omega(x,y)}, \tag{2}$$

where $\mu_\Omega(x,y)$ and $\sigma_\Omega(x,y)$ are the mean and standard deviation of a $N \times N$ neighborhood centered at (x,y). After

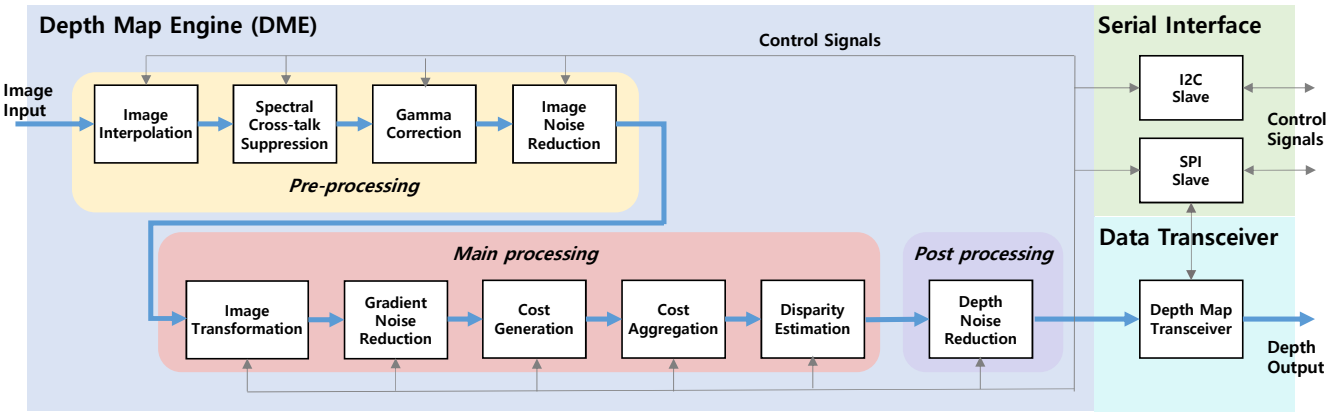


Fig. 3. Top-level hardware architecture of the depth map processor (DMP).

local normalization, mean filtering is performed to remove the noise exaggerated after the gradient operator. Afterwards, sum of absolute difference (SAD) is used for cost-volume generation.

SAD has significant computational advantage over other metrics such as normalized cross correlation (NCC). Local normalization is performed only once over the stereo images. In NCC, however, local normalization needs to be performed for every slice of the cost-volume. Furthermore, SAD-based cost-volume generation requires only simple addition/subtraction whereas NCC requires multiplication for stereo matching.

For cost aggregation, we have used a single path variant of semi-global matching (SGM) [12]. Typically, 16 or 8 paths are used with SGM. However, there are two distinct disadvantages of this approach in terms of hardware complexity. First, the cost volume for the whole image frame needs to be stored as paths are both from top to down and bottom to up. Second, for each path the cost volume over all the disparity levels needs to be stored. For alleviating the excessive memory requirements, we apply the following equation for SGM to compute the aggregated cost.

$$A_r(x,y,d) = C(x,y,d) + min \left\{ \begin{array}{c} A_r(x-1,y,d) \\ A_r(x-1,y,d-1)+P_1 \\ A_r(x-1,y,d+1)+P_1 \\ \min_i A_r(x-1,y,i)+P_2 \end{array} \right\} - \min_k A_r(x-1,y,k), \tag{3}$$

where the last term is included for normalization, $C(x,y,d)$ is the cost at the current pixel (x,y) and disparity d, and P_1 and P_2 are penalty terms. The penalty P_1 is typically a constant and the penalty P_2 is inversely related to the gradient value. The circuit for this implementation had a high iteration bound, limiting the speed of the system. To reduce the iteration bound, we performed aggregation separately at odd and even-numbered pixels. Winner-takes-all is used for disparity estimation at a pixel.

3.3. Post-processing

Numerous methods have been proposed in the past for depth map post-processing, which include weighted median [13], joint-bilateral [14] and joint-guided filtering [15]. These approaches require significant computational power. To reduce the power requirement, we use the following method for depth post-processing.

$$D(x,y) = \frac{\sum_{i,j\in\psi} D_i(i,j) 1\{|I(i,j)-I(x,y)|<S\sqrt{I(x,y)}\}}{\sum_{i,j\in\psi} 1\{|I(i,j)-I(x,y)|<S\sqrt{I(x,y)}\}}, \tag{4}$$

where D_i is the disparity before post-processing, S is a constant, ψ is a square window centered at (x,y) and $1\{.\}$ is the indicator function such that it returns 1 if the condition in the parenthesis is true. In detail, the average depth of inliers is returned by the post-processing method, where the inliers are determined by a threshold set by the Shot-noise assumption over the image. This approach provides a computationally simple, edge-aware approach towards post-processing.

4. EXPERIMENTAL RESULTS

The proposed hardware system was implemented on the Impress FPGA board, which includes Artix-7 xc7a200t FPGA. Also, DDR SDRAM is also available on the board. The DDR SDRAM is used as a frame buffer to display the disparity results on a display. Fig. 5 shows the FPGA board and OA camera for the experiment.

The hardware system is connected externally through the peripheral blocks in Fig. 4. The proposed system can input stereo images from both PC (through USB 3.0) as well as OA camera. An HDMI interface is used to output depth from DMP.

Our experiments show that the DMP can be implemented over a relatively small FPGA such as Artix 7. The proposed

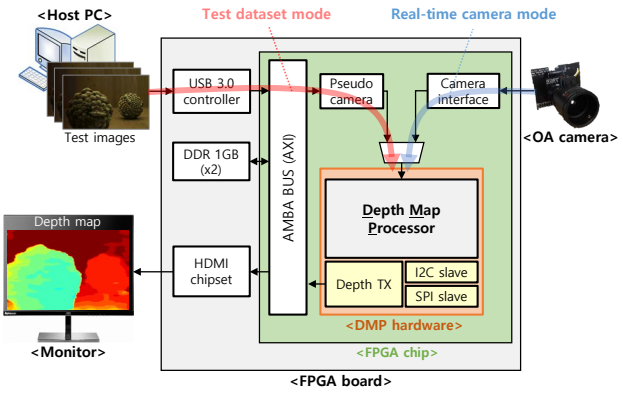


Fig. 4. Block diagram of the evaluation hardware platform

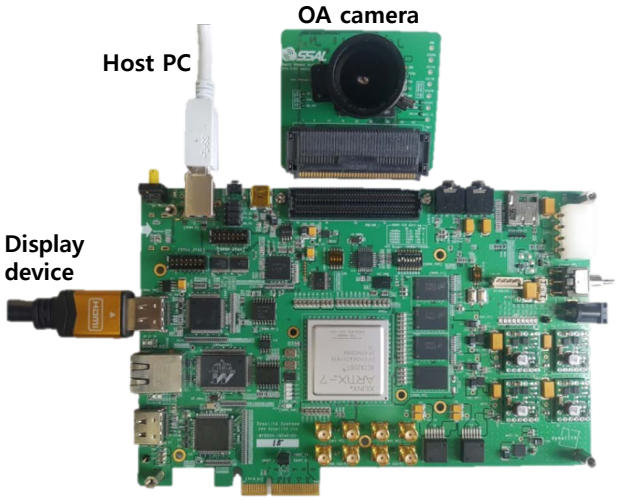
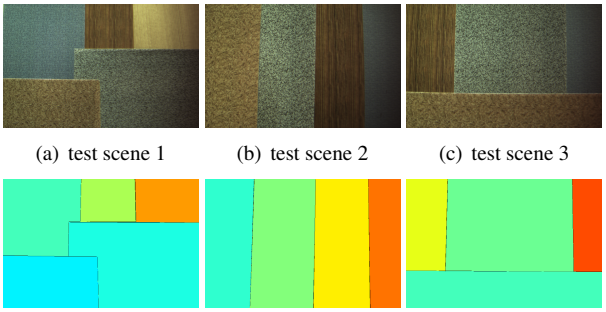


Fig. 5. FPGA board and OA camera

hardware consumes 25% of registers (96,399 / 269,200), 64% of LUTs (86,932 / 134,600), and 54% of memory (25,269bits / 46,200bits) of the available resources on the FPGA. A frame rate of 37.8 fps for VGA (640x480) was achieved while consuming 0.224μJ/pixel.

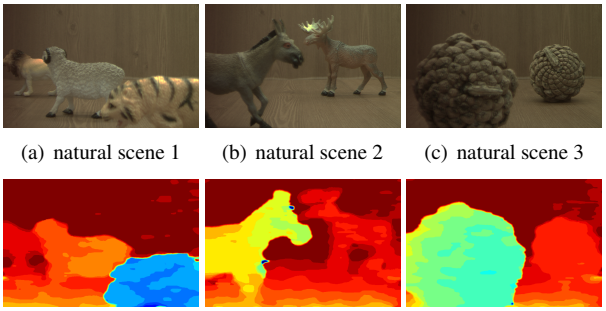
Conventional stereo datasets such as Middlebury [16] and KITTI [17] cannot be used to evaluate the performance of OA-based hardware as these have no IR information. To this end, we developed the ground-truth depth of scenes with flat surfaces, shown in Fig. 6. Object segmentation and plain fitting algorithms were employed to obtain the ground-truth. Percentage of bad classification (PBC) and RMSE were used to evaluate the performance of the proposed DMP. In PBC, an error is declared at each pixel if difference between the obtained disparity and the ground-truth is more than or equal to two. Table 1. shows the result of the accuracy measurement. Three test scenes with the ground-truth are evaluated on the proposed hardware. The number of disparity levels is 31. An average PBC of 2.8% and an average RMSE of 1.27 were obtained by the proposed DMP. Some qualitative results



(a) test scene 1 (b) test scene 2 (c) test scene 3

(d) ground-truth 1 (e) ground-truth 2 (f) ground-truth 3

Fig. 6. Test images and ground-truth. (a) test scene 1, (b) test scene 2, (c) test scene 3, (d)-(f) ground-truth depths of each scenes. Warm color shows far distance.



(a) natural scene 1 (b) natural scene 2 (c) natural scene 3

(d) estimated depth 1 (e) estimated depth 2 (f) estimated depth 3

Fig. 7. Natural scenes and estimated depths of the proposed hardware. (a) natural scene 1, (b) natural scene 2, (c) natural scene 3, (d)-(f) estimated depths of each scenes. Warm color shows far distance.

are shown in Fig. 7.

5. CONCLUSION

This paper proposes OA-based hardware architecture providing depth information with a single shot in real-time. All functional blocks including core depth extraction module and peripheral modules for external communication are integrated within a single FPGA for fast and low-energy depth extraction. The proposed system can be operated at 37.8fps on VGA images while maintaining 2.8% of bad classification.

Acknowledgment

"This work was supported by MSIP as GFP / (CISS-2013M3A6A6073718)"

Table 1. Accuracy measurement of test scenes.

	scene1	scene2	scene3	average
RMSE	1.41	1.03	1.37	1.27
PBC (%)	2.97	2.32	3.12	2.80

6. REFERENCES

[1] Chong-Min Kyung, Muhammad Umar Karim Khan, Asim Khan, Woojin Yun, Young-Gyu Kim, Yeongmin Lee, Jinyeon Lim, Wonseok Choi, and Said Homidov, "Offset aperture: A passive single-lens camera for depth sensing," Submitted to Computer Vision and Pattern Recognition (CVPR), 2017.

[2] Yasufumi Amari and EH Adelson, "Single-eye range estimation by using displaced apertures with color filters," in *Industrial Electronics, Control, Instrumentation, and Automation, 1992. Power Electronics and Motion Control., Proceedings of the 1992 International Conference on*. IEEE, 1992, pp. 1588–1592.

[3] Yosuke Bando, Bing-Yu Chen, and Tomoyuki Nishita, "Extracting depth and matte using a color-filtered aperture," in *ACM Transactions on Graphics (TOG)*. ACM, 2008, vol. 27, p. 134.

[4] Eunsung Lee, Wonseok Kang, Sangjin Kim, and Joonki Paik, "Color shift model-based image enhancement for digital multifocusing based on a multiple color-filter aperture camera," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 2, 2010.

[5] Vladimir Paramonov, Ivan Panchenko, Victor Bucha, Andrey Drogolyub, and Sergey Zagoruyko, "Depth camera based on color-coded aperture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–9.

[6] Seunghun Jin, Junguk Cho, Xuan Dai Pham, Kyoung Mu Lee, Sung-Kee Park, Munsang Kim, and Jae Wook Jeon, "Fpga design and implementation of a real-time stereo vision system," *IEEE transactions on circuits and systems for video technology*, vol. 20, no. 1, pp. 15–26, 2010.

[7] Paolo Zicari, Stefania Perri, Pasquale Corsonello, and Giuseppe Cocorullo, "Low-cost fpga stereo vision system for real time disparity maps calculation," *Microprocessors and Microsystems*, vol. 36, no. 4, pp. 281–288, 2012.

[8] Christian Banz, Sebastian Hesselbarth, Holger Flatt, Holger Blume, and Peter Pirsch, "Real-time stereo vision system using semi-global matching disparity estimation: Architecture and fpga-implementation," in *Embedded Computer Systems (SAMOS), 2010 International Conference on*. IEEE, 2010, pp. 93–101.

[9] Michael Kuhn, Stephan Moser, Oliver Isler, Frank K Gurkaynak, Andreas Burg, Norbert Felber, Hubert Kaeslin, and Wolfgang Fichtner, "Efficient asic implementation of a real-time depth mapping stereo vision system," in *Circuits and Systems, 2003 IEEE 46th Midwest Symposium on*. IEEE, 2003, vol. 3, pp. 1478–1481.

[10] John Iselin Woodfill, Gaile Gordon, and Ron Buck, "Tyzx deepsea high speed stereo vision system," in *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*. IEEE, 2004, pp. 41–41.

[11] Graham D. Finlayson, Steven D. Hordley, and Paul M. Hubel, "Color by correlation: A simple, unifying framework for color constancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1209–1221, 2001.

[12] Heiko Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 2, pp. 807–814.

[13] S-J Ko and Yong Hoon Lee, "Center weighted median filters and their applications to image enhancement," *IEEE transactions on circuits and systems*, vol. 38, no. 9, pp. 984–993, 1991.

[14] Carlo Tomasi and Roberto Manduchi, "Bilateral filtering for gray and color images," in *Computer Vision, 1998. Sixth International Conference on*. IEEE, 1998, pp. 839–846.

[15] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," in *European conference on computer vision*. Springer, 2010, pp. 1–14.

[16] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition*. Springer, 2014, pp. 31–42.

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3354–3361.