Full HD Real-time Depth Estimation Algorithm and Hardware Implementation

Hejian Li, Ping An
Key Laboratory of Advanced Display and System
Application Ministry of Education,
Shanghai, China
Leehejian@163.com

Guowei Teng, Yifan Zuo, Zhaoyang Zhang
School of Communication and Information Engineering,
Shanghai University
Shanghai, China
Anping@shu.edu.cn

Abstract—**This paper presents a novel implementation for real-time depth estimation which can support full HD resolution for 3D video stream. We propose a package of solution to address the major difficulties in depth estimation, such as expensive computation, depth accuracy, real-time implementation and high definition progressing. Different from the typical single stereo vision methods, we take fusion strategy which can reduce errors caused by individual measure. The FPGA-based design is selected for real-time implementation which can achieve the maximum processing speed of 125fps, with maximum disparity search range of 240 pixels, full HD (1920 × 1080p) resolution image. Multi-resolution method which includes interpolation method and synthesizers are used for improve the efficiency and stability. The implementation can be included in video system for live 3DTV application and can be used as an independent hardware module in low-power application.**

***Keywords-** Depth estimation; full HD; hardware implementation; Real-time*

I. INTRODUCTION

With the development of stereo technology, variety of 3D applications appears such as 3-D movies, 3-D Blue-ray, and the auto-stereoscopic display appears for commercial 3DTV application. These promote 3D video acquisition and production technology research and promote the stereoscopic 3D to reach high standard now[1]. But technically still has some room for improvement and there are some unresolved central technology. The high-definition real-time depth estimation and efficient video codec requires further research and development. In this paper we focus on the difficulties of depth estimation, such as boundary reserving of depth map, improvement of processing speed resulted by algorithm complexity.

In communication system, depth estimation is generally performed at the transmit end, obtaining the depth map, and at the receive end, adopts depth image based rendering (DIBR) to generate arbitrary virtual view[2]. There has been lots of research on depth estimation, the mode of implementation has software, hardware, and Scharstein et al.[3] summarize the algorithm which include global, local area based. Woodfill et al.[5] present the well known Deep sea stereo vision system which speed is 200fps at 512×480 resolution. But it is about 23 fps if only 1920×1080 is considered. Riechert et al.[7] present the software algorithm which is capable of processing 1080p disparity maps in real-time, but on a system with two general purpose CPU's and two high-end GPU's which not suitable for embed application.

The proposed method and its FPGA implementation can achieve the maximum processing speed of 125fps, with maximum disparity search range of 240 pixels, full HD (1920 × 1080p) resolution image. The proposed system includes video capture, format conversion, stereo matching, post-processing and disparity visualization steps to achieve depth map in real-time. For efficient matching algorithm, referring both to ensure the accuracy of the estimated disparity, but also consider the convenience for hardware implementation, this paper adopts fusion local matching strategy. The method includes multi-resolution operation for supporting full HD resolution video input, and using synchronous design to overcome the instability problem introduced by clock domain crossing (CDC) operation. Finally, the experimental evaluations demonstrate the effectiveness and efficiency of our implementation. .

The remainder of this paper is structured as follows. Section Ⅱ interprets the algorithms and flow. Section Ⅲ summarizes the hardware implementation. Results are discussed in Section Ⅳ. Conclusion shows in section Ⅴ.

II. DEPTH ESTIMATION ALGORITHM

The matching algorithm adopts fusing matching strategy that the combined image matching measure successfully reduces the errors caused by individual measures. The local algorithms offer a good potential for parallelization and is well suited for hardware implementation due to its high degree of parallelization. The task of a stereo vision algorithm is to analyze the images captured by a pair of cameras and to extract the objects shift in both images. This shift is counted in pixels and called disparity d. According to the geometry constrain, it can obtain the real-world depth $Z=bf/d$, where b, f are the baseline and focal length of camera pair. The flow of proposed depth estimation algorithm, include pre-processing, image matching, post-processing etc. Dashed box means optional.

The proposed area-based matching method considers both the retention of edges; also considers the full utilization of the texture region sensitive. It is a combination algorithm of Census Transform (CT) and sum of Absolute Difference (SAD), which attempt to exploit the complementary advantages of both approaches. CT approach has advantages in bias-independent and homogenous area and low hardware complexity, while SAD has advantage in feature-rich areas. Both algorithms combine to form a complementary enhancement. Cost aggregation is used to further improve the distinguish degree of matching cost.

A. Census Transform

In early 90's 19century, Woodfill [8] proposed Census transform, which belongs to non-parametric local transforms. Compare conventional algorithm, it can avoid noise between image pairs and simplify the hardware design by integral calculation. In particular, matching performance is high in the structural feature highlighted regions, such as area near object boundaries. One CT value of current pixel is bits array which summarizes local image structure generated in a specified window. CT transform values of left can be expressed by the following formulas and CT of right view has the same expression.

$$I'_r(u,v) = \otimes_{n\in N} \otimes_{m\in M} \xi(r(u,v), r(u+n, v+m)) \qquad (1)$$

$$\xi(p_1, p_2) = \begin{cases} 0, p_1 > p_2 \\ 1, p_1 \le p_2 \end{cases}$$

where *I'r, I'l* are the transform value of right and left view, corresponding to the pixel at coordinates (*u*,*v*) position. The operator $\otimes$ denotes a bit-wise catenation and M×N the census window size. ξ is 0 when pixel original value *p1* is bigger than *p2*, otherwise it is 1.

B. Fusion Matching Algorithm

Stereo matching is a progress of finding corresponding pixels in image pairs. In the area-based algorithms, epipolar constrain is used to decrease computational complexity. Matching costs of fusing algorithm were used for best matching point searching within the searching range of disparity. Fig.1 shows the matching processing which along the disparity D axis. It shows that the disparity searching process is to find optimized point in different disparity plane.

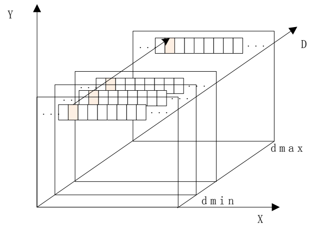

Fig. 1. Initial matching cost values of one image in different disparity planes

The initial matching cost calculation includes two parts, respectively, hamming distance obtained from census transformed image and SAD value based on the original image. The hamming distance between two pixels can be calculated by the following equation (2). The equation (3) shows the cost calculation processing. *I'* and *I* represent the CT value and intensity value of pixel respectively. *(u,v)* is the coordinate of current pixel.

$$C_{CT}(u,v,d) = Ham\min g(I'_1(u,v), I'_2(u+d,v)) \qquad (2)$$

$$C_{\mathrm{SAD}}(\mu,v,d) = \sum_{(i,j)\in w} \left| I_1(u+i, v+j) - I_2(u+i+d, v+j) \right| \qquad (3)$$

Equation (4) shows the combination of the equation above C_{CT} and C_{SAD} under the same block size *w*, then get the primitive matching cost *C(u,v,d)* corresponding to one pixel in the disparity plane showed in Fig. 2.

$$\begin{aligned} &\mathrm{C}(\mu,v,d) = \rho_{SAD} \sum_{(i,j)\in w} \left| I_1(u+i,v+j) - I_2(u+i+d,v+j) \right| + \\ &\rho_{CT} Ham\min g(\otimes_{i,j\in w} \xi(I_1(u,v), I_1(u+i,v+j)), \qquad (4) \\ &\otimes_{i,j\in w} \xi(I_2(u+d,v), I_2(u+d+i,v+j))) \end{aligned}$$

where ρ_{SAD}, ρ_{CT} present the integration parameters. This combines the results of the previous mentioned to achieve the total value of the primitive matching cost. As can be seen from the formula, the same block is used multiple times. This offers the possibility to optimize memory in hardware design. After the computation of the absolute differences, the primitive matching costs are aggregated over a block of pixels to improve accuracy. In this work, we use static block sizes. The costs aggregation is assumed that neighboring pixels, except at disparity discontinuities, have a similar disparity level, so a costs aggregation makes the matches more unique. The larger the block size used, the larger the impreciseness at object borders and the more time consuming in processing. In order to keep the good trade-off between quality and processing time, aggregation with 5x5 or 9x9 template generate new costs value for the pixel in center of the pattern.

C. Interpolation Method for Multi-resolution

High resolution offers more details compared to low resolution images, and High resolution is the trend for future applications. For hardware implementation, the main difficult is high resolution processing faces current resource limitation, on the one hand because of the higher data bandwidth, and the data cache capacity constraints. The paper adopts multi-resolution processing to achieve efficient hardware design for a high resolution image, the appropriate interpolation method is used for depth estimation. As one key step for multi-resolution process, interpolation includes Nearest Neighbor Interpolation, Linear interpolation, B-Spline Interpolation, Joint bilateral filter up-sampling and so on. Since joint bilateral filtering up-sampling not only requires the use of low-resolution image data, but also need to use additional guidance image for separately calculating spatial filter kernel and range filter kernel, this will result in additional storage consumption and higher complexity in hardware implementation. The up-sampling in the paper will not consider using joint bilateral filter, rather trying other methods described earlier, and choose Nearest Neighbor Interpolation of considering the complexity and the result effect. Equation (5) is the general expression of interpolation.

$$f(\mathbf{x}) = \sum_{k\in Z^q}^{n} c_k \cdot \beta(\mathbf{x}-\mathbf{k}) \ \forall \mathbf{x} \in R^q \qquad (5)$$

where interpolated value $f(\mathbf{x})$ at some coordinate $\mathbf{x}$ in a space of dimension *q* (*q* = 2 for images) is calculated as the coefficients C_k from the known samples f_k and interpolation sample weights $\beta(\mathbf{x}-\mathbf{k})$ which is basis function.

Nearest Neighbor Interpolation provides a constant-repeated input pattern. The main advantage of this interpolation is its simplicity. The advantage of this procedure is that the whole interpolation process is completely reversible and there is no loss of the initial data in the final image.

D. Other Progressing

In addition to the above mentioned matching core algorithm and multi-resolution operation, other operations are used in different parts of the system to enhance the performance of depth estimation, such as the costs aggregation for improving the accuracy, and post-processing is processing to reduce various artifacts introduced in the depth map, such as unaligned edges of object, occlusion errors, image noise etc. The left-right consistency check(LR-check) occludes the points where the two images are not negatives of each other. Sub-pixel processing can improve accuracy in sub-pixel level. Filter can reduce the image noise.

III. HARDWARE IMPLEMENTATIONEASE OF USE

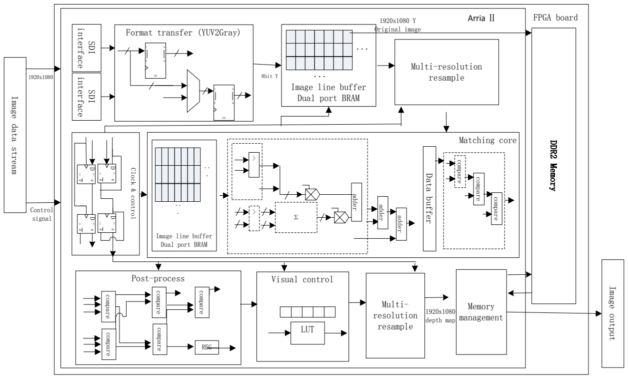

Fig. 2. Block diagram of our depth estimation hardware platform

The structure of the overall hardware architecture is summarized in the Fig. 2. The processed image data stream from the binocular cameras with SDI interface are converted from YUV values to 8-bit grayscale intensities and fed to the following module. One of the main components is matching core to search for correspondence, which uses the combination algorithm of CT and SAD. Subsequently, the disparity calculation is performed, discarding possible ambiguities. In addition to the core module, format transfer, multi-resolution resample module, clock control, post-process, visual controller, synchronizer designed to avoid CDC problems and various buffers and memory controller modules are components of it. The memory controller interfaces to external DDR2 memory and BRAMs which construct the line buffers.

To verify the depth estimation module described above, the FPGA-based PCB board is setup, as shown in Fig. 3. The function includes two channel video acquisitions, real-time depth estimation, and depth map can be transmitted through the PCI interface to the server for further processing, including encoding, visual view generation. All the function above has been implemented in our lab.

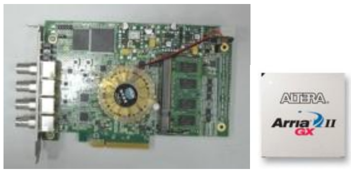

Fig. 3. PCB board with Altera FPGA chip

IV. RESULTS

The proposed algorithm implementation is on Altera FPGA chip, the type is EP2AGX260EF29C4. Module design uses Verilog hardware design language (HDL). This section will first summarize the implementation result and resource consumption of design, and then analyze the quality of the depth map, and finally through power analysis and data processing calculation to further analyze the performance of the whole implementation.

TABLE I. SPECIFICATIONS OF TARGET DEVICE FOR ALGORITHM IMPLEMENTATION(EP2AGX260EF29C4)

Combinational ALUTs	25,510 / 205,200 (12 %)
Memory ALUTs	24 / 102,600 (< 1 %)
Dedicated logic registers	41,780 / 205,200 (20 %)
Total registers	42216
Total pins	206 / 432 (48 %)
Total virtual pins	16
Total block memory bits	1,511,888 / 8,755,200 (17 %)
DSP block 18-bit elements	0 / 736 (0 %)
Total GXB Receiver Channel PCS	8 / 12 (67 %)
Total GXB Receiver Channel PMA	8 / 12 (67 %)
Total GXB Transmitter Channel PCS	8 / 12 (67 %)
Total GXB Transmitter Channel PMA	8 / 12 (67 %)

The performance of the implementation can reach the highest resolution in 1920×1080px, a frame rate of 125 frame per second, disparity search range is 240 pixels. Module implement on Altera FPGA. The resource consumption is shown in Table I, the internal storage consumption mainly come from the CT transform, the matching cost aggregation, as well as the filter window operations which need row buffers for parallel calculation. So the storage units increase with the increase of resolution and the processing window height. Fortunately, from the results effect, it is not the bigger the window handle better [9], so take into account the quality assurance and resource consumption to determine the appropriate processing window, such as CT transform can choose between 9 ×9-17 × 17. The core module combined with the external cache processing part and a clock management unit.

In order to verify the effect of the output of the proposed algorithm, mainly is the quality of the depth map. Depth maps captured from the implementation system show the effect in real scene. Original image and depth map are obtained through the implementation of the system in our lab environment. Fig. 4. shows original left view is on the left, right is the corresponding depth map, and below is pseudo-image of depth map. It shows that the object edges is relatively clear, rich texture area has high accuracy. Low texture area needs to be further improved. Depth map is not directly to watch, mainly applied in DIBR. From the perspective of rendering, low texture area has similar value in video image, thus has high error tolerance in rendering processing. Relatively the quality of rich texture region in depth map determine the quality of rendered image. Our method keeps the advantage of accurate rendering. The color bar in Fig. 4. shows the Color and distance relationship.

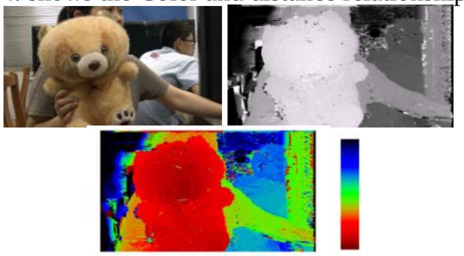

Fig. 4. Result of implementation of proposed depth estimation algorithm

TABLE II. FPGA DEVICE POWER DISSIPATION CHARACTERISTICS(EP2AGX260EF29C4)

Total Thermal Power Dissipation	3130.79 mW
Core Dynamic Thermal Power Dissipation	1279.04 mW
Core Static Thermal Power Dissipation	1439.44 mW
I/O Thermal Power Dissipation	412.31 mW

The power consumption has been a matter of concern, especially under equipment miniaturization trend, it is becoming a critical issue. Table II shows the power consumption of proposed implementation. With respect to hundreds of Watts power CPU, our approach is about 3Watt, it has obvious advantages and it is suitable for portable application.

TABLE III. HARDWARE PERFORMANCE COMPARISON

Method	**Performance**				
	Device	*Resolution*	*Algorithm*	*fps(disp)*	*Mde/s*
[5] Woodfill	ASIC	512×480	Census	200(52)	2556(49)
[6] Hadjitheophanous	DSP	640×480	Rank/SGM	30(128)	1179(9.2)
[4] Jin	FPGA	640×480	Census	230(32)	2261(70.6)
[7] Riechert	PC GPU	1920×1080	SGM	16	(33.2)
Proposed method	FPGA	1920×1080	Census/SAD	125(240)	62208(259.2)

In addition to evaluate the whole performance, Table III gives the comparison to previous works. In depth estimation system, usually the higher the resolution, the faster the processing speed requirements, the larger the disparity range, it will need critical requirements in terms of hardware storage, operation frequency and logic resource usage and it will provide large data throughout. So Mde/s=width×height×disps×fps/1000000 is used for performance evaluation. This is more meaningful in line with the overall performance of the system. Table 3 lists the comparison of this method with other solutions. For our proposed system, the resolution is full HD, the maximum disparity search range is 240 pixels which can support search closer object than method with low disparity search range. From the Mde/s index also can be seen that the performance improvement. Compared with the previous design, the whole performance of our method is boosted dramatically.

V. CONCLUSION

In this paper, we proposed fusion matching algorithm and hardware implementation for full HD real-time depth estimation application. The speed is up to 125fps for 1920×1080pixel image resolution pair with 240 disparity levels. The proposed fusion algorithm is used to enhance the matching accuracy since it reduce errors caused by individual algorithm and it has high paralleled processing potential which is benefit for hardware implementation. FPGA-based design is used for implementation. Through the performance analysis of various aspects of the design, in addition to the highest specification, the resource consumption and power consumption are analyzed. The whole performance of Mde/s is improved dramatically. In future work, more efficient interpolation method will be used to improve the quality of depth map.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China, under Grants U1301257, 61172096 and 60902085, the key Project of Shanghai Science and Technology Commission, under Grant 12DZ2293500.

REFERENCES

[1] P. Merkle, et al., “3D video: Acquisition, coding, and display,” Proceedings of the 2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE), pp. 127-128, 2010.

[2] Karsten Muller, et al., “3-D Video Representation Using Depth Maps,” Proceedings of the IEEE, 2011, 4(99):643-656.

[3] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two frame stereo correspondence algorithms,” International Journal of Computer Vision, Springer, vol. 47, no. 1-3, pp. 7-42, 2002.

[4] Seunghun Jin, et al., “FPGA Design and Implementation of a Real-Time Stereo Vision System,” IEEE Transactions on Circuits and Systems for Video Technology,20(1):15-26, 2010.

[5] John Iseling Woodfill, et al., The Tyzx DeepSea G2 vision system, a taskable, embedded stereo camera, in: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshops.

[6] S. Hadjitheophanous, et al., “Real-time stereo vision system using semi-global matching disparity estimation: Architecture and FPGA-implementation,” Embedded Computer Systems (SAMOS), 2010 International Conference on, 93 - 101

[7] C. Riechert, et al., “Real-time disparity estimation using line-wise hybrid recursive matching and cross-bilateral median up-sampling,” in Proceedings of the IAPR 21st International Conference on Pattern Recognition, pp. 3168-3171, 2012

[8] Zabih R,Woodfill J. Nonparametric local transforms for computing visual correspondence. Proceeding of European Conference on Computer Vision, Stockholm, Sweden. 1994: 151–158.

[9] Hadhoud, M.M., " New trends in high resolution image processing", Photonics and Its Application, 2004. The Fourth Workshop on, P2 -23