Optical Engineering

OpticalEngineering.SPIEDigitalLibrary.org

Full high-definition real-time depth estimation for three-dimensional video system

Hejian Li
Ping An
Zhaoyang Zhang

SPIE.

Full high-definition real-time depth estimation for three-dimensional video system

Hejian Li,[a,b] Ping An,[a,b,*] and Zhaoyang Zhang[a,b]
[a]Key Laboratory of Advanced Display and System Application, Ministry of Education, Shanghai 200072, China
[b]Shanghai University, School of Communication and Information Engineering, Shanghai 200072, China

Abstract. Three-dimensional (3-D) video brings people strong visual perspective experience, but also introduces large data and complexity processing problems. The depth estimation algorithm is especially complex and it is an obstacle for real-time system implementation. Meanwhile, high-resolution depth maps are necessary to provide a good image quality on autostereoscopic displays which deliver stereo content without the need for 3-D glasses. This paper presents a hardware implementation of a full high-definition (HD) depth estimation system that is capable of processing full HD resolution images with a maximum processing speed of 125 fps and a disparity search range of 240 pixels. The proposed field-programmable gate array (FPGA)-based architecture implements a fusion strategy matching algorithm for efficiency design. The system performs with high efficiency and stability by using a full pipeline design, multiresolution processing, synchronizers which avoid clock domain crossing problems, efficient memory management, etc. The implementation can be included in the video systems for live 3-D television applications and can be used as an independent hardware module in low-power integrated applications. [DOI: 10.1117/1.OE.53.9.093108]

Keywords: depth estimation; real time; full high definition; three-dimensional video.

Paper 141040 received Jul. 1, 2014; revised manuscript received Sep. 1, 2014; accepted for publication Sep. 3, 2014; published online Sep. 24, 2014.

1 Introduction

By using dense depth information, three-dimensional (3-D) video systems[1] such as 3-D Blue-ray, 3-D television (TV) sets provide stereo video, which has multiple views for different viewers. The results of depth estimation can also be used for driving assistance for automatic driving vehicles and robot navigation. In 3-D TV and free-viewpoint TV (FTV)[2] applications, depth image based rendering (DIBR) [3] is used to generate the arbitrary virtual views. The partial content of each view is synthesized into one image which includes eight or more views for autostereoscope display. So resolution reduction of each viewing zone is inevitable for multiview 3-D. High definition is necessary to provide high image quality in this condition.

Depth estimation has been thoroughly studied and a broad spectrum of approaches are summarized and compared by Scharstein and Szeliski.[4] Hardware design based on a digital signal processor,[5] application-specific integrated circuit (ASIC),[6,7] or field-programmable gate array (FPGA) is more competent for real-time depth estimation implementation and design embedding than PC-based design. State-of-the-art implementations continually improve the performance. Jin et al.[8] present an FPGA-based real-time stereo vision system which processes 640 × 480 images with a disparity range of 32 pixels in 230 fps. Considering only the pixel throughput, this system would be adequate for 1920 × 1080 with 34 fps. Riechert et al.[9] present a software algorithm that is capable of processing 1920 × 1080 disparity maps in real time, but on a system with two general-purpose CPUs and two high-end graphic processing units (GPUs), which is not suitable for embedded application. The software algorithm presented by Mei et al.[10] presents a system with good performance in accuracy, and currently, it is the second performer in the Middlebury benchmark.[11] But it is implemented on a PC with a GPU card and gets a low-resolution sequence (e.g., 512 × 384, 60 disparity levels) in ∼10 fps.

Researches implement and evaluate variable efficient matching algorithms in order to get a high-performance depth estimation based on hardware design. Scharstein and Szeliski's taxonomy[4] has summarized the stereo matching algorithm. Compared to global algorithms of stereo matching (e.g., graph cut),[12] belief propagation (BP),[13] whose calculation is a complex local (area-based) algorithm, has hardware friendly characteristics. A single matching algorithm strategy, such as the sum of absolute difference (SAD) based[14–17] or census based[18–21] were widely used in the hardware designs, early research, and other matching methods, such as the dynamic programming algorithm[22] and BP,[23] were implemented based on a GPU platform. Previous works illustrate that a hardware solution provides real-time processing. But it is inevitable that the single local method introduces some errors under some conditions. For example, the census transform (CT) method introduces errors in the repeat edges. Low image resolution is another problem that exists in available implementations. With the increasing requirements for accuracy and image size, there are two strategies that appear in hardware implementation. One is the semi-global matching (SGM) method. The other is a fusion algorithm, which involves more than one algorithm to take advantage of different methods in the implementation.

Considering the analysis above, a high-definition depth estimation (HDDE) system, which is a real-time FPGA implementation, is proposed in this paper. It is capable of

*Address all correspondence to: Ping An, E-mail: anping@shu.edu.cn processing full HD (1920 × 1080) content with a maximum processing speed of 125 fps and a maximum disparity search range of 240 pixels. The HD depth map enhances the quality of the stereo content and the large depth range ensures that objects close to the cameras can be measured. The fusion matching strategy is used and implemented. The method includes a multiresolution operation for supporting a full HD resolution video input and using a synchronous design to overcome the instability problem introduced by the clock domain crossing (CDC) operation. Finally, in order to evaluate the performance of the design, the source usage condition and power consumption are analyzed, and the mega disparity evaluation per second (MdeS) is used to illustrate the overall performance.

The remainder of this paper is structured as follows. Section 2 analyzes the algorithms of the depth estimation, which include the fusing strategy stereo matching algorithm and a multiclock domains operation. Section 3 summarizes the details of hardware implementation for the HDDE system. Results of implementation are discussed in Sec. 4. A conclusion is shown in Sec. 5.

2 Depth Estimation

In a stereo vision system, image matching algorithms are important. The task of a stereo vision algorithm is to analyze the images captured by a pair of cameras and to extract the object shift in both images. This shift is counted in pixels and is called disparity *d*. According to the geometry constraints, the real-world depth is $Z = bf/d$, where b and f are the baseline and focal length of the camera pair. The flow of the proposed depth estimation includes preprocessing, image matching, postprocessing, etc. This section describes the algorithm of the fusing matching method by adopting the idea that the combined image matching measure successfully reduces the errors caused by individual measures. The FPGA implementation is used because it is suited for consumer applications in terms of size, cost, and power consumption. This is one main motivation to implement this algorithm.

2.1 *Stereo Matching*

Based on the requirements of the hardware design, the fusion strategy algorithms adopted in the implementation should have mutually reinforcing features and have the potential of parallel processing. SAD and CT belong to local (area-based) algorithm methods. Table 1 illustrates a comparison of CT and SAD with different characteristics. The CT approach has advantages in being bias-independent and having a homogenous area and low hardware complexity, while SAD has an advantage in feature-rich areas, especially in texture repeated regions. They are strongly complementary to each other and, hence, have huge potential for fusion in matching efficiency.

Table 1 Comparison of census transform and sum of absolute difference (SAD).

Criterion	Census	SAD
Bias-independent	**Yes**	No
Homogenous area	**Strong**	Weak
Feature-rich areas	Weak	**Strong**
Hardware complexity	**Low**	High

The initial matching cost calculation includes two parts, respectively, the Hamming distance obtained from the census transformed image and the SAD value based on the original image. The hybrid matching costs[10] can be expressed by Eq. (1):

$$C(P,d) = \rho[C_{\text{Census}}(P,d), \lambda_{\text{census}}] + \rho[C_{\text{SAD}}(P,d), \lambda_{\text{AD}}], \tag{1}$$

where λ_{AD} and λ_{Census} present the integration parameters, which can be adjusted to control the influence of outliers. Two individual cost values of C_{Census} and C_{SAD} are computed. $\rho(C,\lambda)$ can be $1 - \exp[-(c/\lambda)]$. With this normalization, Eq. (1) will not be severely biased by one of the measures.

CT is a nonparametric local transform proposed by Woodfill[24,18] in the early '90s of the 20th century. Compared to a conventional algorithm, it can avoid noise between image pairs introduced by different cameras and simplify the hardware design with an integral calculation. In particular, the matching performance is high in the structural feature highlighted regions, e.g., areas near object boundaries. One CT value of a current pixel is the bits array, which summarizes the local image structure in a specified window. The CT value of left/right view I' is equal to $\otimes_{n\in N}\otimes_{m\in M}\,\xi[p(u,v), p(u+n, v+m)]$, where the operator $\otimes$ denotes a bitwise catenation, $M \times N$ is the mesh window size, and $\xi(p_1, p_2)$ is 1 when the pixel original value p_1 is bigger than p_2, otherwise it is 0. (u, v) are the coordinate values of the corresponding pixel. If k is the number of the bit of each CT value, k is equal to $w \times w - 1$ bit as the window size is $w \times w$. Figures 1(a) and 1(b) show an example of CT where the mesh window is 3×3. The Teddy image and its CT image are shown in Figs. 1(c) and 1(d).

Different window sizes result in transform values with different lengths, which impacts the matching results. Meanwhile, the computation complexity increases with the increasing window size. Figures 2(a) and 2(b), respectively, show the correct percentages of disparity estimation and the time consumption based on a double core 2.67 GHz PC with different window sizes from 7×7 to 37×37. They shows that with an increasing transform window, the accurate rate of the depth map does not increase linearly, while the computation obviously increases. Therefore, a tradeoff is needed between the window size and computation complexity in the hardware design.

After running CT, the Hamming distance, $\text{Hamming}[I'_1(u,v), I'_2(u+d,v)]$, is used as one part of the initial matching cost. I'_1 and I'_2 are CT values. The value of the Hamming distance is the sum of the bitwise exclusive OR of a pixel pair.

Another famous local matching method is SAD.[25] Its cost function C_{SAD} is the absolute difference of the pixels' intensity values,

$$C_{\text{SAD}}(\mu,\nu,d) = \sum_{(i,j)\in w} |I_1(u+i, v+j) - I_2(u+i+d, v+j)|. \tag{2}$$

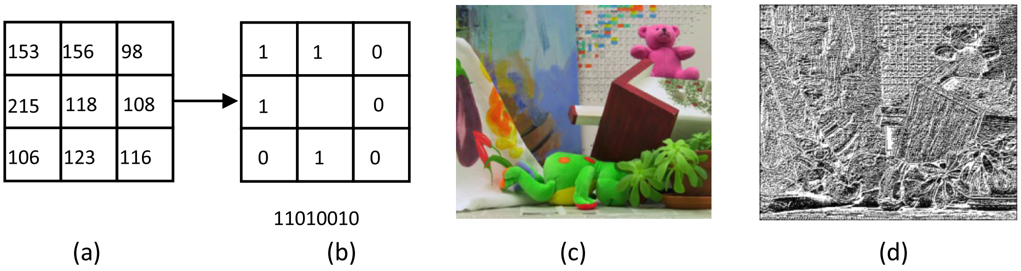

Fig. 1 Census transform example: (a) 3 × 3 census mask window, (b) census transform result of center pixel, (c) Teddy, (d) census transform results of Teddy.

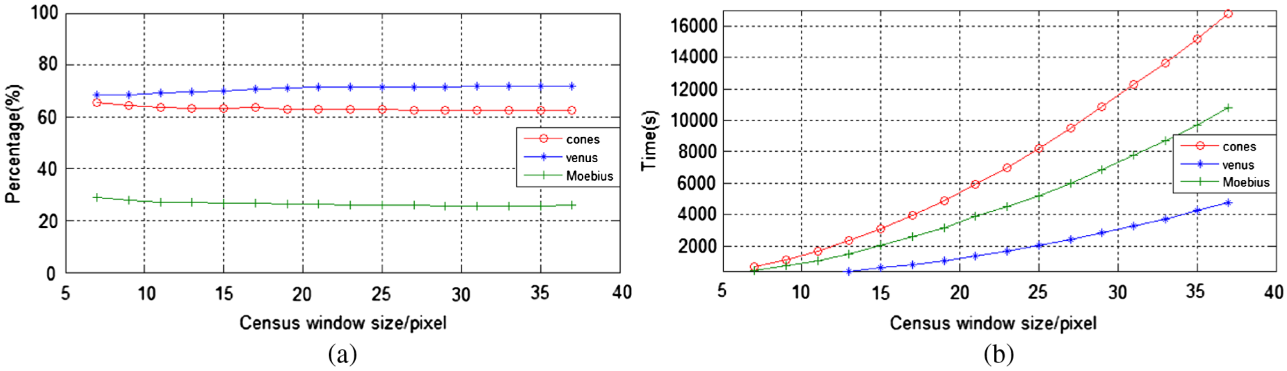

Fig. 2 Matching quality and time consumption for different census mask sizes.

I_1 is the intensity value of the primary stereo image. I_2 is the intensity value of the secondary stereo image at disparity level d. The coordinates are $(u + i,\ v + j)$.

In the fusion steps, different combinations of the two matching methods can be used. The normalization method and other methods are used to control the proper weight of the basic matching costs. Following efficient combination, matching costs are provided for the disparity optimization processing.

2.2 *Multiscale Processing*

High-resolution image processing is one of the trends in 3-D video[26] and it is necessary to provide a high-quality image in an autostereoscopic display. In contrast to natural video signals, depth maps are characterized by piecewise smooth regions bounded by sharp edges. A smooth interior surface is a benefit for a multiscale operation. We find that the depth map can obtain a higher quality in an interpolation process compared to a normal texture image. Figure 3 shows the interpolation quality difference between the depth map and the texture image with different interpolation methods. The interpolation methods of 1 to 7 represent, respectively, nearest, bilinear, bicubic, box, lanczos2, lanczos3, and spline interpolation.[27] This shows that no matter what interpolation is used, the depth map has a better performance than the texture image. Since joint bilateral filtering[27] upsampling not only requires using the low-resolution image data but also needs an additional guidance image for separately calculating the spatial filter kernel and the range filter kernel, it will

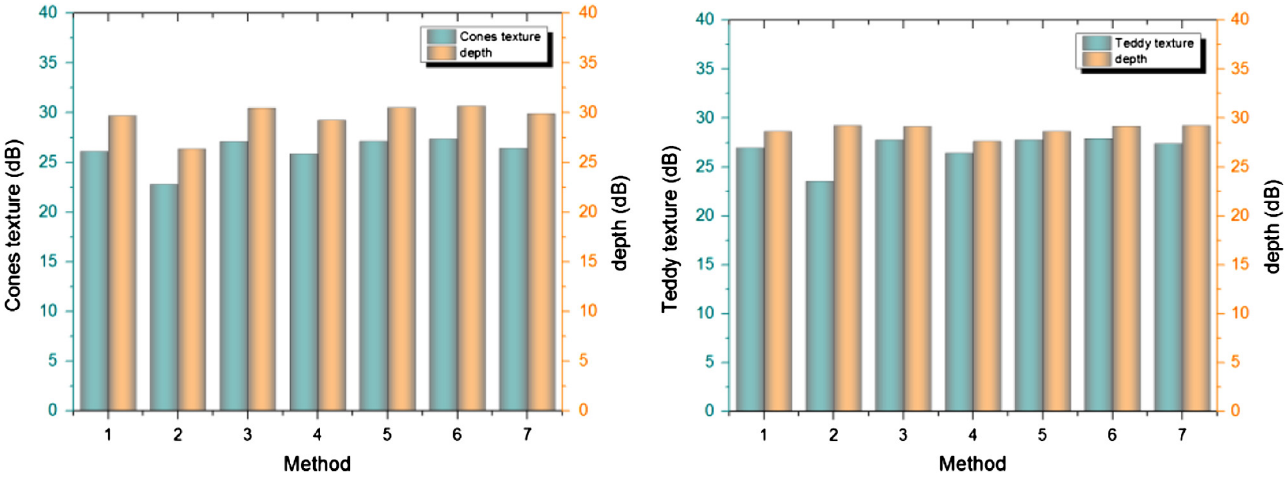

Fig. 3 Interpolation performance for depth map and texture image.

result in additional storage consumption and high complexity. Therefore, the joint bilateral filter will not be used for upsampling in this paper. The method that can be used in this design depends not only on the quality of the results, but also on the implementation complexity. Nearest interpolation is adopted in this design and synthesizers are used in the hardware design to avoid meat-stability problems since the multiclock domains are involved in different resolution operations.

In addition to the above-mentioned image matching algorithm and multiscale operation, refining processes are also used to enhance the performance of the depth estimation. The improved cross-based regions are processed for efficient cost aggregation. Support regions based on cross-skeletons allow fast aggregation with middle-ranking disparity results. As an optimizer, winner-takes-all (WTA) is used for the initial disparity map. A refining process to correct various disparity errors is used for improving the disparity results. The left-right consistency check occludes the points where the two disparity maps are not negatives of each other.[8] Subpixel disparity[8] is another postprocessing method, which can use a parabolic fitting to generate the disparity in the subpixel accuracy. Filtering can be used as a fitting to achieve subpixel-level image filtering. These are used to resolve problems such as misleading occlusions and not being aligned with objects and outliers.

3 Hardware Implementation

The top-level block diagram of the proposed depth estimation hardware architecture is summarized in Fig. 4. The HDDE system involves eight submodules and internal block random access memorys (BRAMs) and an external double data rate 2 (DDR2) memory for data buffering. The processed source image data streams are captured from the professional binocular cameras with a serial digital interface and the data format is converted from YUV to 8-bit gray-scale intensities. When the depth estimation is completed, the depth maps according with the original texture image are sent out for further processing such as coding. BRAMs are used to store rows of image data for the window operation in the algorithm implementation. The external memory bandwidth is important for the depth estimation of high-resolution images. The proposed memory management module controls the bandwidth and the data allocation scheme based on the asynchronous first input first output (FIFO) architecture. The eight submodules consist of three main submodules and five submodules.

1. Matching core: This module mainly realizes the matching algorithm, which includes CT, Hamming distance, fusion of CT and SAD, etc. A register transfer level (RTL) description is used for logic synthesis. It receives pixelwise image data and exports pixelwise disparity data.
2. Multiresolution processing module: It realizes upsampling or downsampling for multiresolution operation in multiple clock domains. Synchronizers are used to resolve CDC problems.
3. Memory management: It controls the data transmission between the internal module and external DDR2 memory. An asynchronous FIFO is used to synchronize the data stream and modulate the data width for data transform efficiency.
4. Other modules: Besides the above main module, input/output interface, format conversion, postprocessing, clock control module, and a visual controller module

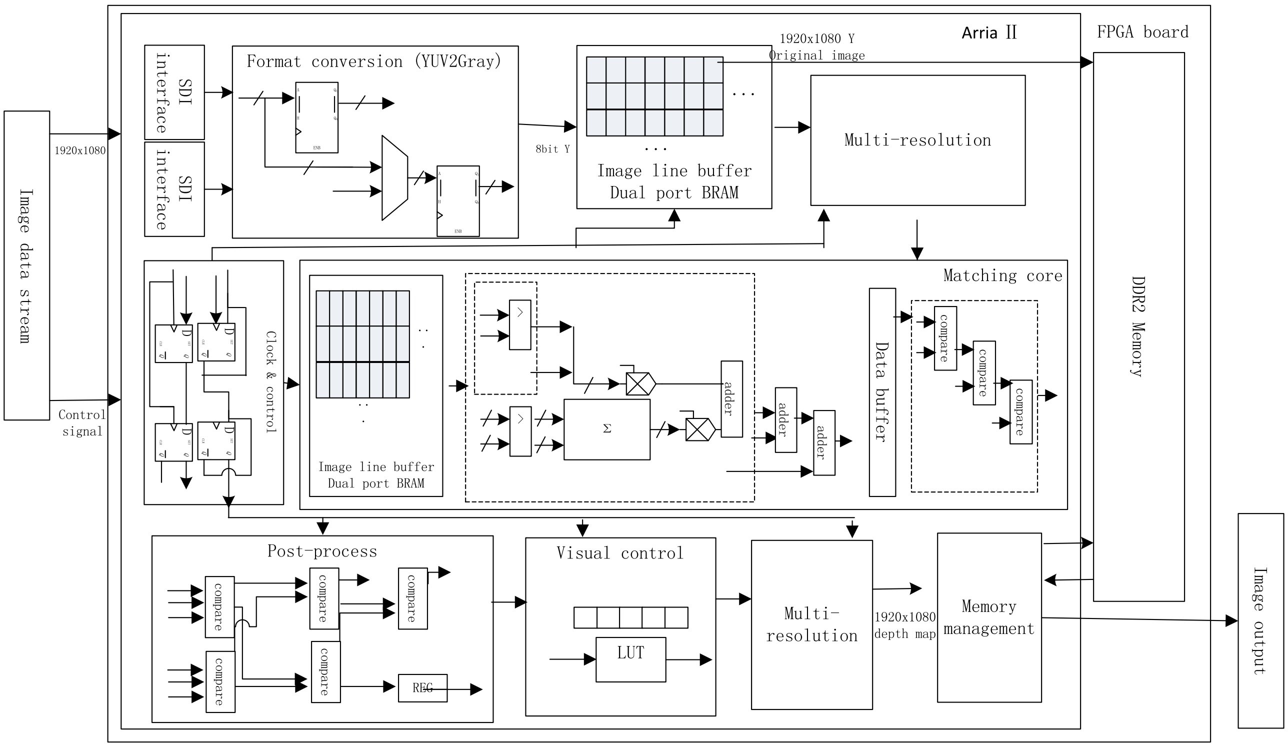

Fig. 4 Hardware architecture block diagram of high definition depth estimation (HDDE) system.

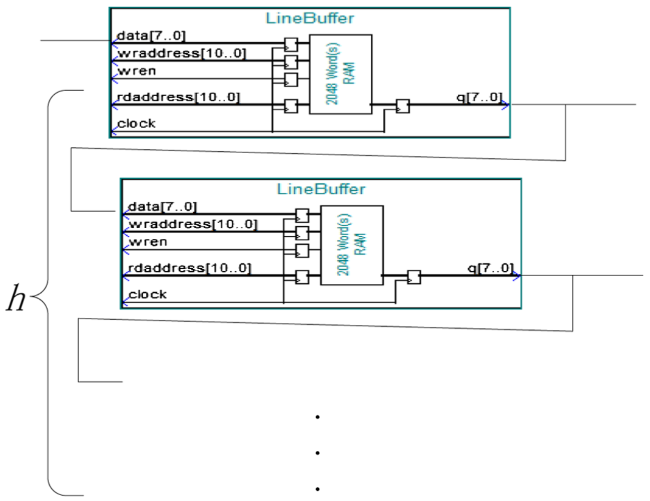

Fig. 5 Line buffers for $w \times h$ window operation.

are involved in the system. These modules provide functions such as data control, data transform, further processing, etc.

3.1 *Matching Core Implementation*

The matching core is the key architecture of the proposed hardware fusion algorithm implementation, which adopts a parallel design to increase the throughput. Calculation of the value of the CT for each pixel is to compare the center pixel with pixels around it in the transform window. To meet this design requirement, it is necessary to set the line buffers with memory depth equal to the window height and the word width of the buffer equal to 8 bits. Meanwhile, the progress should be finished in one pixel clock to ensure real-time processing. The Altera intellectual property (IP) core of a dual-port RAM was used to realize the line buffers for the windows operation. The size of every line buffer is determined by the size of the input image. As shown in Fig. 5, a RAM with read and write ports composes the line buffers for further implementation. The number of line buffers is determined by the window height h and the width of line buffer is determined by the image horizontal resolution. In the experiments, the full HD images (1920×1080 pixels) were progressed, so nine line buffers are used for 9×9 windows and the minimum memory depth of each line buffer is 1920 words, where the word width is 8 bits. The outputs of each line buffer are merged together by using register arrays to form the data for the windows operation, as shown in Fig. 6. The diagram gives an example of a 5×5 window operation, which is composed of $w \times h$ cascade registers. The CT diagram in Fig. 6 is used to get the CT value of one pixel in serial. One CT value is 80 bits for a 9×9 window and is saved in the memory unit with an 80-bit word width. Figure 7 shows the processing architecture of the CT, which consists of comparators and outputs bit arrays with $m = w \times h - 1$ bits. As the output of this submodule, the original image data and CT value will be fed to next stage.

The initial matching cost value of each pixel is calculated by mixing the Hamming distance of the CT value and the SAD value determined by the original intensity value. In the matching processing, the current pixel and all pixels included in the disparity range in the reference image take part in the matching cost calculation, which generates $Ln = d\,\max - d\,\min + 1$ matching cost values for each pixel. To ensure real-time processing, the output of all the Ln matching costs of the current pixel should be completed simultaneously. Parallel processing of the costs for each pixel is achieved through cascade register units. Hardware architecture of the Hamming distance module and the SAD consists of adder and comparator logic. Figure 8 shows the cost fusion calculation architecture. The parameters were utilized in fusing the CT and SAD costs. Figure 9 shows the logic design of the Hamming distance, which uses a fine grain pipeline method. The example shows the comparison of two 16-bit CT values by three-level pipeline structures.

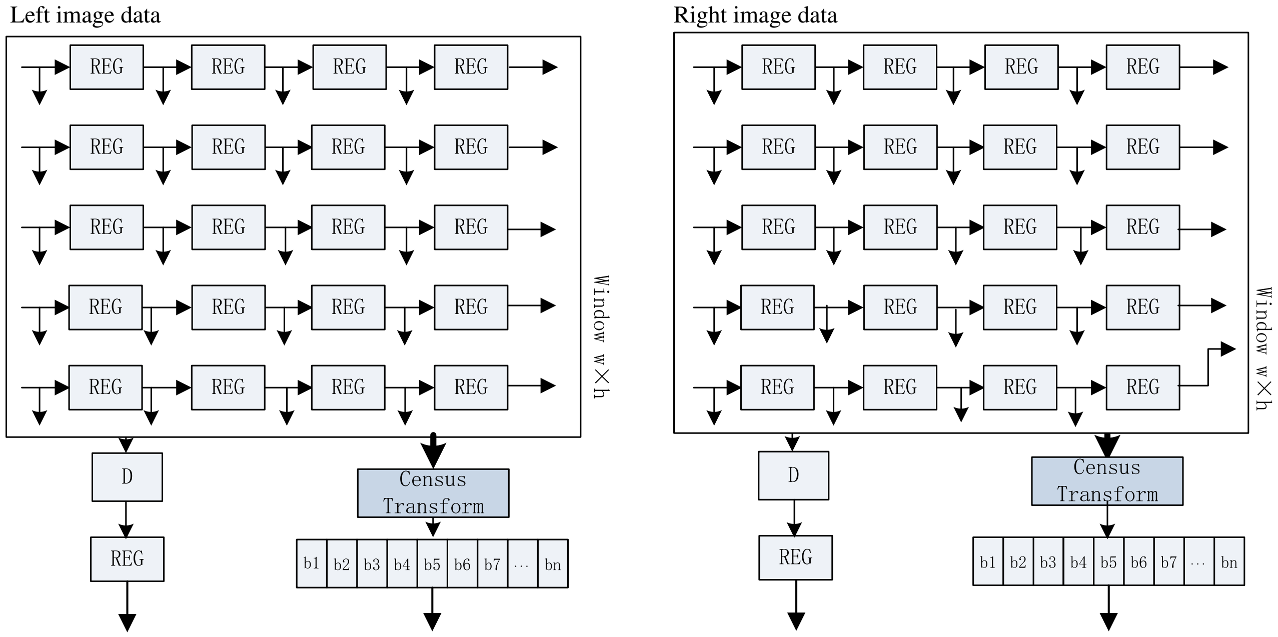

Fig. 6 Register arrays for window operation.

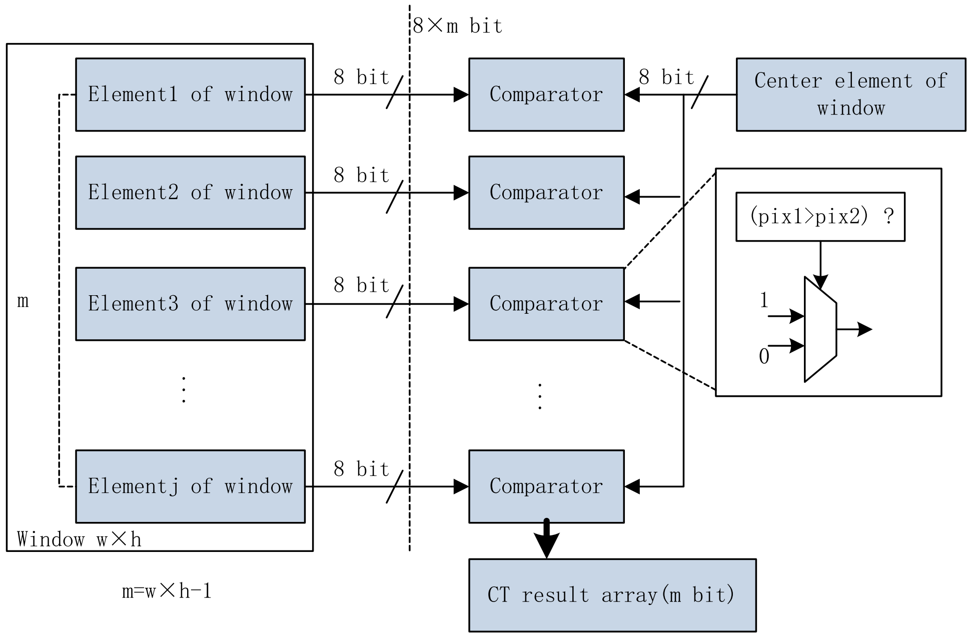

Fig. 7 Block diagram of census transform.

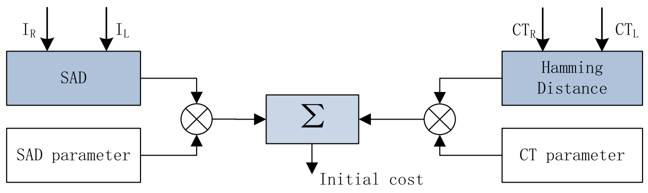

Fig. 8 Census transform and sum of absolute difference calculation for the initial matching cost.

To further improve the matching accuracy, the cost aggregation in the specified window is carried out in each of the disparity plans. In order to keep a good trade-off between the quality and processing time, a 5×5 squared window aggregation is used. The windows operation in aggregation processing is the same as the method in CT, which uses the line buffers and register arrays. The five-stage pipeline is adopted. The first stage outputs are 10 bits, and the second stage result is 11 bits, so a 14-bit result is output in the final stage. The cost results are used to get the disparity using optimization processing.

3.2 *Multiclock Domain Design*

The design has to include several different clock domains because the FPGA chip and peripheral devices have different operating frequencies; another reason is the multiscale operation which has been used to improve the processing efficiency of the high-resolution video. When data are transferred from one clock domain to another, the received data risk including errors because of the existence of metastability. So the synchronizer is necessary for stable data communication. Based on the metastability characteristics of a synchronizing flip-flop, synchronizer reliability is typically expressed in terms of the mean time between failures

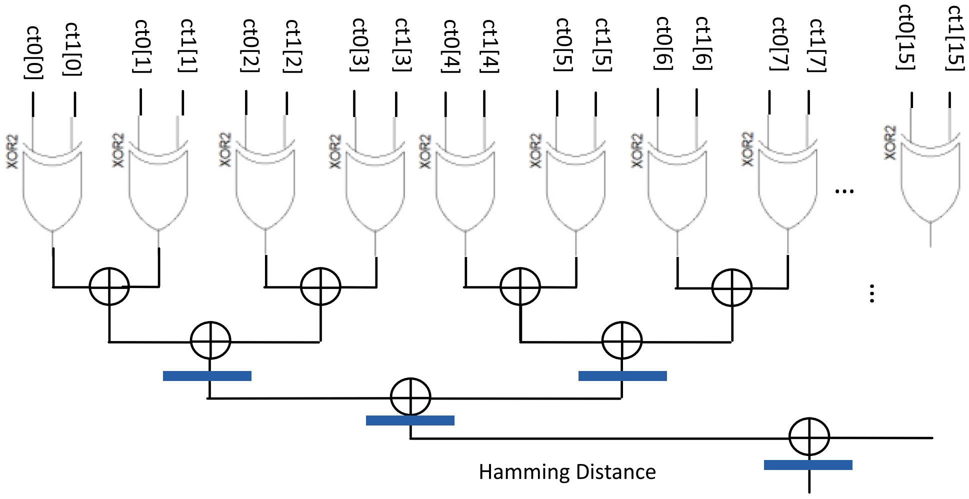

Fig. 9 Logic design of Hamming distance.

(MTBF).[28] It can be calculated by $\text{MTBF} = (e^{T_{\rm r}/\tau}/T_0 f_{\rm s} f_{\rm d})$, where τ is the settling time of the flop, T_W is a parameter related to its time window of susceptibility, $f_{\rm s}$ is the synchronizer's clock frequency, and $f_{\rm d}$ is the frequency of pushing data across the clock domain boundary. In the proposed design, synchronizers are used in the multiscale operation and DDR2 memory management to improve the system stability. Besides the synchronizer above, asynchronous FIFO is suitable for data bus synchronization. It is a typical structure for asynchronous FIFO.[29] In the application of DDR2 write management of the HDDE system, a data buffer consists of the FIFO memory unit and the FIFO control module which generates the control signals.

3.3 *Memory Organization*

High-resolution processing and complex algorithm processing need mass memory. So memory management and organization are important parts of an HDDE system. The external DDR2 memory is arranged for the frame buffer, and BRAM is used for the slice storage. In order to realize the communication between FPGA internal data and DDR, a data input/output buffer module is set up to resolve the CDC problem. As shown in Fig. 10, write control in the internal clock domain, memory buffer, and read control in DDR2 clock domain compose the DDR2 data input buffer. In addition, the data bit width has been adjusted to make efficient use of the DDR2 data transfer ability. The data that were read under read control were transferred to DDR2 through a direct memory access module. The DDR access flow is shown in Fig. 11. As mentioned above, there are similar modules that are responsible for inverse data access.

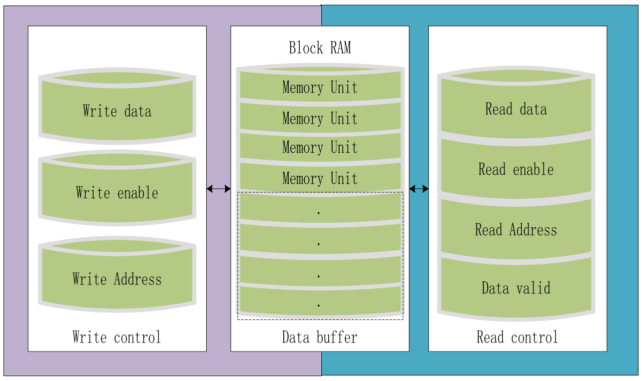

Fig. 10 Data input buffer module scheme.

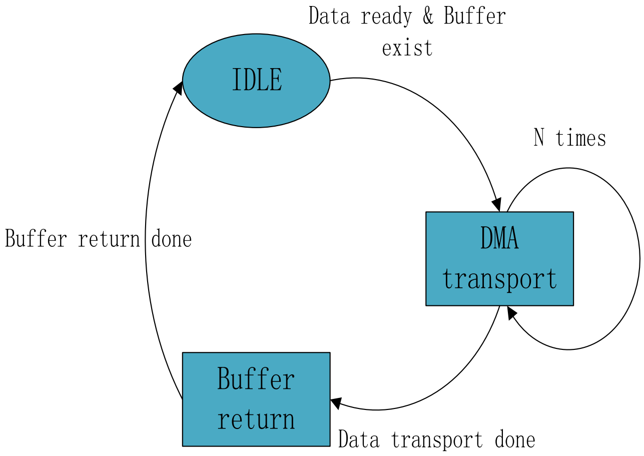

Fig. 11 Double data rate 2 memory access flow.

3.4 *Other Submodules*

Other submodules, such as the input signal transform, WTA, postprocessing, or visual quantification module, are implemented using logic gates and logic units, such as comparators, counters, multiplexers, and finite-state machines. Figure 12 shows the architecture of the basic unit of a WTA module, which includes a comparator and multiplexer. Figure 13 shows the design of the median filter which can reduce the random noise introduced in the disparity map assignment stage.

An Optional modules are the module of disparity converted to depth and depth quantization. If the camera pair is fixed, the camera intrinsic and external parameters are constant and the depth in the real world is in fixed inverse proportion to the disparity. The processing of depth quantization can be expressed as $I_{\rm d}(z) = \text{round}\{255[(1/z) - (1/z_{\max})]/[(1/z_{\min}) - (1/z_{\max})]\}$. Depth representation has two major advantages. First, because depth is the distance between the object and the camera, it avoids the disparity dependence on camera parameters, which is favorable for free viewpoint rendering based on a depth map. On the other hand, the converted quantization results in that object close to the camera has a fine resolution and an object far from the camera has a coarse resolution, which satisfies the human vision character. Since directly using the formula to design modules is a high resource consumption, a look-up table is used to realize the conversion processing.

4 Results and Discussion

The depth estimation system we proposed, the HDDE system is implemented and the stereo vision system based on it is setup for performance evaluation. The binocular professional cameras of Panasonic AG-3DA1MC are used to provide the full HD 1920 × 1080 pixels video streams. An autostereoscopic display with an eight-view display LCD is used for effect evaluation. The HDDE system implemented on an EP2AGX260EF29C4 of the Arria II GX

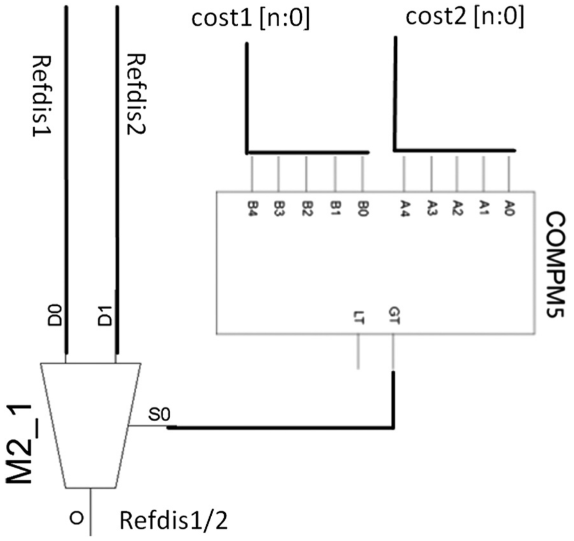

Fig. 12 Logic unit of winner-takes-all.

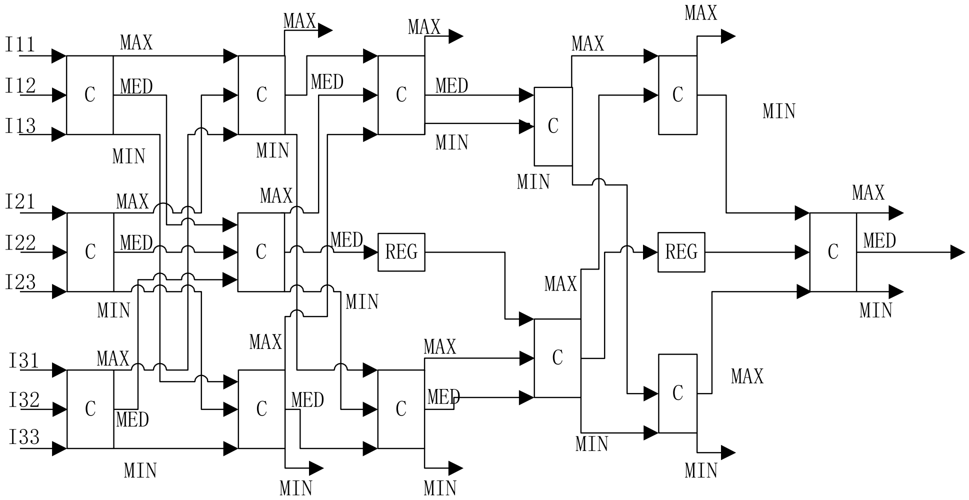

Fig. 13 Circuit scheme of median filter.

family is bounded on the print circuit board (PCB) board with a peripheral component interconnect (PCI) interface. Following the real-time depth map generation, a server PC is used for the remaining processing (e.g., DIBR). Part of the system is shown in Fig. 14, which also shows the FPGA-based PCB board.

In the experimental video system, the depth map can be directly displayed on the LCD TV or can be used to generate a virtual image based on the DIBR. The autostereoscopic display shows the real-time stereo video based on the original images and synthesized virtual images. The visible depth map and stereo video effect are utilized to provide a direct measure of the depth estimation performance.

The HDDE system is capable of processing full HD content with a processing speed of 125 fps and a disparity search range of 240 pixels. This section will summarize the implementation results and resource consumption of the design, analyze the depth map quality, and, finally, further analyze the performance of the whole system through power analysis and data processing calculations. The core module of the HDDE system mentioned below refers to design modules and does not include peripheral interface IPs and management modules. The implemented system includes the core module, the necessary interface IP module, the external memory management, etc.

4.1 *Hardware Implementation Result*

The HDDE system has been implemented using Verilog, simulated using Mentor Graphics ModelSim 6.6 and

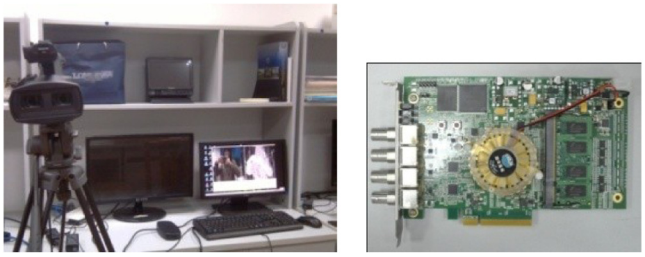

Fig. 14 Stereo vision system and field-programmable gate array based PCB board for HDDE system implementation.

Table 2 Specifications of target device for core module and high definition depth estimation (HDDE) system implementation on EP2AGX260EF29C4 of Arria II GX.

Resource	Core	HDDE implemented system
Combinational adaptive look up tables (ALUTs)	15,779/205,200 (8%)	25,510/205,200 (12%)
Memory ALUTs	86/102,600 (<1%)	24/102,600 (<1%)
Dedicated logic registers	21,500/205,200 (10%)	41,780/205,200 (20%)
Total registers	21,500	42,216
Total pins	55/432 (13%)	206/432 (48%)
Total virtual pins	0	16
Total block memory bits	1,070,092/ 8,755,200 (12%)	1,511,888/ 8,755,200 (17%)
Total gigabit transceive block (GXB) receiver channel physical coding sublayer (PCS)		8/12 (67%)
Total GXB receiver channel physical medium attachment (PMA)		8/12 (67%)
Total GXB transmitter channel PCS		8/12 (67%)
Total GXB transmitter channel PMA		8/12 (67%)
Total PLLs		2/6 (33%)
Total delay–locked loops		1/2 (50%)

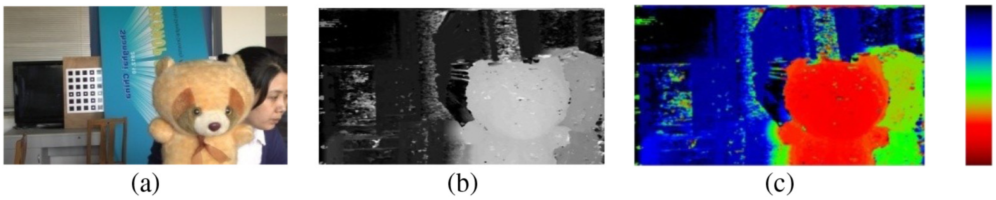

Fig. 15 Results captured in the implemented video system. (a) Left input image of a real-scene scenario. (b) Depth map. (c) False-color result rendered for optical viewing, including the color bar.

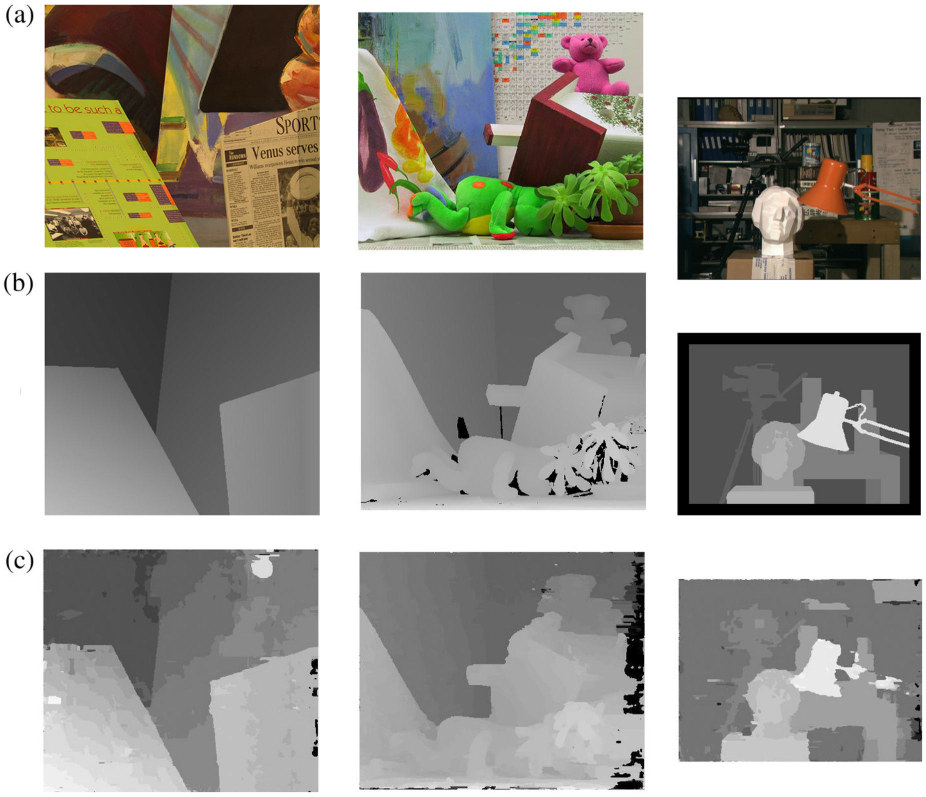

Fig. 16 Matching results. (a) Original Middlebury stereo data set. (b) Ground truth. (c) Matching results.

executed using Quartus II tool. Except for the design, a core module has been developed to allow an easy migration to other FPGA or ASIC technologies. The resource consumptions of the core module and HDDE system implementation are shown in Table 2. The internal storage consumption mainly comes from the CT and the matching cost aggregation, as well as the filter window operations which need to setup line buffers for parallel calculation. So the storage unit will increase with the increasing resolution and the processing window size. Taking into account the quality assurance and resource consumption, we can determine the appropriate size of the window, e.g., between 9×9 and 17×17 can be chosen for CT. Compared to the core module, the logic unit and the BRAM of the HDDE system are slightly increased because of the additional external communication interface IP resource. The phase locked loops (PLLs) are used for setting multiple clock domains.

4.2 *Depth Map Analysis*

In order to evaluate the effect of the depth estimation results, depth maps of real scenes collected directly from the implementation system are used for evaluation. Figure 15(a) is captured in the laboratory environment, Fig. 15(b) is the corresponding depth map, and Fig. 15(c) is the false-color rendered image of the depth map. To help optical viewing, the distance coordinate in the real world is also shown in the figure. As can be seen, there is high accuracy in texture-rich regions which are fit for real scene reconstruction. This exhibits the characteristics that not only the regions with repeating texture but also the area of objects' edges are all concerned and also the reliability of the algorithm implementation is illustrated.

We use data sets of Middlebury for our algorithm performance measurement; the images Venus, Teddy, and Tsukuba are measured. Figure 16 shows the results and the last line is the disparity image obtained by the proposed method, which also includes the postprocessing mentioned above. Disparity maps have high accuracy in high texture regions, which is the advantage of the SAD algorithm. Good edge information illustrates the characteristics of the CT method.

4.3 *Further Analysis of the Performance*

A hardware-based design, especially a semiconductor design, can achieve low power costs in the application

Table 3 Field-programmable gate array (FPGA) device power dissipation characteristics.

	Core	HDDE implemented system
Total thermal power dissipation	801.07 mW	3130.79 mW
Core static thermal power dissipation	789.71 mW	1279.04 mW
Input/output thermal power dissipation	11.36 mW	1439.44 mW

system. The power consumption is a critical matter of concern in hardware design, especially under the equipment miniaturization trend. Altera PowerPlay Power Analyzer is used to make a power consumption analysis of our HDDE system. Table 3 shows the results of power consumption. Compared to the typical power of an Intel Core I7 processor, the working condition is ∼338 W[30] and the execution HDDE system power consumption is ∼3.2 W. The power consumption of the core module is only 801 mW. Reducing the resource usage rate of FPGA or reducing the operation clock frequency can further reduce the power consumption.

One single aspect, e.g., image resolution, is not sufficient to measure the performance of the hardware implementation of depth estimation. The performance of the system can be illustrated by its data processing capabilities. Therefore, this paper uses megapixels per second and mega disparity evaluation per second (MdeS) as the assessment criteria. MdeS can be described as MdeS $=$ width $*$ height $*$ disps $*$ fps$/$1,000,000. It is more meaningful in line with the overall performance of the system. Table 4 compares the existing approaches with the hardware implementations mentioned in this paper. It can be seen that the maximum frame rate in Refs. 8 and 30 is high, but in the 3-D TV, real-time 3-D video applications only need $>$30 fps for National Television Standards Committee (NTSC) and 25 fps for Phase Alternating Line (PAL). Considering only the pixel throughput, they would be adequate for 1080 pixels with 34 and 65 fps because their resolution is VGA. The proposed implementation can compete with other hardware architectures in terms of full HD depth maps in real time and can dramatically boosts the overall performance.

5 Conclusions

An FPGA architecture of a real-time depth estimation system, the HDDE system, has been proposed and evaluated in this paper. It is capable of processing full HD (1920 × 1080) resolution stereo video at 125 fps with 240 disparity levels. It boosts the overall performance of the stereo vision system by efficient hardware parallel implementation, including pipeline and block window data parallel architectures. It uses a fusion stereo matching strategy, which enhances the quality of the depth map with hybrid CT and SAD matching algorithms. An autostereoscopic

Table 4 Disparity estimation performance comparisons.

Reference	Performance					
	Resolution	Frame rate (fps)	Megapixels per second	Mega disparity evaluation per second (disparity range)	Method	Processor type
Presented work	1920 × 1080	125	259.2	34,937 (240)	Census+SAD	FPGA
24	320 × 240	42	3.2	77.4 (24)	Census	Multi FPGA
19	512 × 480	200	49.2	2555.9 (52)	Census	Application-specific integrated circuit
31	320 × 240	100	7.7	245.8 (32)	SAD	FPGA
25	512 × 512	25.6	6.7	1711.3 (255)	SAD	FPGA
32	384 × 288	50	5.5	88.5 (16)	SAD	Digital signal processor (DSP)
23	320 × 240	16	1.2	19.6 (16)	Belief propagation	Graphic processing unit (GPU)
33	640 × 480	320	98.3	7864.3 (80)	SAD	FPGA
34	640 × 480	30	9.2	1179.6 (128)	Rank+SGM	FPGA
35	640 × 480	75	23.0	1152 (50)	Census	PC/DSP/GPU
36	320 × 240	75	5.8	144 (25)	SAD	FPGA
8	640 × 480	230	70.7	2261 (32)	Census	FPGA
30	640 × 480	440	135.2	10,813 (80)	SAD	FPGA
21	640 × 480	60	18.4	3538.9 (192)	Census	FPGA
37	640 × 480	30	9.2	589.8 (64)	Census+SGM	Multi FPGA

display with an integrated renderer can be connected to the hardware and provides a high-quality multiview video. The system performs with high efficiency and stability by using a full pipeline design, multiresolution processing, synchronizers which avoid CDC problems, efficient memory management, etc. Results of power analysis demonstrate that the proposed architecture is capable of a low-power application. In future work, we intend to search for a more robust postprocessing method to further improve the quality of the depth map and try to use it as a part of the 3-D video system for high-performance 3-D reconstruction.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants U1301257, 61172096, and 61171084.

References

1. K. Muller, P. Merkle, and T. Wiegand, "3-D video representation using depth maps," *Proc. IEEE* **99**(4), 643–656 (2011).
2. M. Tanimoto, "Overview of free viewpoint television," *Signal Process.: Image Commun.* **21**(6), 454–461 (2006).
3. M. Tanimoto et al., "Reference software for depth estimation and view synthesis," MPEG/M15377, Archamps, France (2008).
4. D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.* **47**(1), 7–42 (2002).
5. C. Cuadrado et al., "Real-time stereo vision processing system in an FPGA," in *Proc. of the IEEE 32nd Annual Conf. on Industrial Electronics*, pp. 3455–3460, IEEE, Paris, France (2006).
6. M. Kuhn et al., "Efficient ASIC implementation of a real-time depth mapping stereo vision system," in *Proc. of the 46th IEEE Int. Midwest Symp. on Circuits and Systems*, pp. 1478–1481, IEEE, Cairo, Egypt (2004).
7. K. Ambrosch and W. Kubinger, "Accurate hardware-based stereo vision," *Comput. Vis. Image Underst.* **114**(11), 1303–1316 (2010).
8. S. Jin et al., "FPGA design and implementation of a real-time stereo vision system," *IEEE Trans. Circuits Syst. for Video Technol.* **20**(1), 15–26 (2010).
9. C. Riechert et al., "Real-time disparity estimation using line-wise hybrid recursive matching and cross-bilateral median up-sampling," in *Proc. of the ICPR 21st Int. Conf. on Pattern Recognition*, pp. 3168–3171, IEEE, Tsukuba, Japan (2012).
10. X. Mei et al., "On building an accurate stereo matching system on graphics hardware," in *Proc. IEEE Int. Conf. on Computer Vision Workshops*, pp. 467–474, IEEE, Barcelona, Spain (2011).
11. http://vision.middlebury.edu/stereo/.
12. Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(11), 1222–1239 (2001).
13. J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(7), 787–800 (2003).
14. L. Mingxiang and J. Yunde, "Stereo vision system on programmable chip (SVSoC) for small robot navigation," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 1359–1365, IEEE, Beijing, China (2006).
15. K. Ambrosch et al., "Hardware implementation of an SAD based stereo vision algorithm," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–6 (2007).
16. J. Ding et al., "Improved real-time correlation-based FPGA stereo vision system," in *Int. Conf. on Mechatronics and Automation*, pp. 104–108 (2010).
17. N. Isakova, S. Basak, and A. C. Sonmez, "FPGA design and implementation of a real-time stereo vision system," in *Int. Symp. on Innovations in Intelligent Systems and Applications*, pp. 1–5, IEEE, Trabzon, Turkey (2012).
18. J. Woodfill and B. Von Herzen, "Real-time stereo vision on the PARTS reconfigurable computer," in *Proc. of the 5th IEEE Symp. on FPGA-Based Custom Computing Machines*, p. 201–210, IEEE, Napa Valley, California (1997).
19. J. I. Woodfill, G. Gordon, and R. Buck, "Tyzx DeepSea high speed stereo vision system," in *Proc. of the 2004 Conf. on Computer Vision and Pattern Recognition Workshops*, p. 41, IEEE, Washington, DC (2004).
20. A. Gardel et al., "Parametric dense stereovision implementation on a system-on chip (SoC)," *Sensors (Basel)* **12**(2), 1863–1884 (2012).
21. J. G. Kim et al., "A real-time virtual re-convergence hardware platform," *J. Semicond. Technol. Sci.* **12**(2), 127 (2012).
22. R. Kalarot, J. Morris, and G. Gimel'farb, "Performance analysis of multi-resolution symmetric dynamic programming stereo on GPU," in *25th Int. Conf. of Image and Vision Computing New Zealand*, pp. 1–7 (2010).
23. Q. Yang et al., "Real-time global stereo matching using hierarchical belief propagation," in *Proc. of the 2006 British Machine Vision Association Conf.*, pp. 989–998 (2006).
24. R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondance," in *Proc. Third European Conf. on Computer Vision*, pp. 151–158, Springer Berlin Heidelberg (1994).
25. S. Perri et al., "SAD-based stereo matching circuit for FPGAs," in *Proc. of the 13th IEEE Int. Conf. on Electronics, Circuits and Systems*, pp. 846–849 (2006).
26. M. M. Hadhoud, "New trends in high resolution image processing," in *Fourth Workshop on Photonics and Its Application*, pp. 2–23 (2004).
27. J. Kopf et al., "Joint bilateral upsampling," *ACM Trans. Graph.* **26**(3), 96 (2007).
28. C. Dike and E. Burton, "Miller and noise effects in a synchronizing flip-flop," *IEEE J. Solid-State Circuits* **34**(6), 849–855 (1999).
29. L. Luo et al., "Design and verification of multi-clock domain synchronizers," in *Int. Conf. on Intelligent System Design and Engineering Application*, Vol. 1, pp. 544–547 (2010).
30. C. Georgoulas and I. Andreadis, "A real-time fuzzy hardware structure for disparity map computation," *J. Real-Time Image Process.* **6**(4), 257–273 (2011).
31. A. Naoulou et al., "An alternative to sequential architectures to improve the processing time of passive stereovision algorithms," in *Proc. of the Int. Conf. on Field Programmable Logic and Applications*, pp. 1–4, IEEE, Madrid, Spain (2006).
32. N. Chang et al., "Real-time DSP implementation on local stereo matching," in *Proc. of the IEEE Int. Conf. on Multimedia and Expo*, pp. 2090–2093 (2007).
33. C. Georgoulas and I. Andreadis, "FPGA based disparity map computation with vergence control," *Microprocess. Microsyst.* **34**(7–8), 259–273 (2010).
34. C. Banz et al., "Real-time stereo vision system using semi-global matching disparity estimation: architecture and FPGA-implementation," in *Int. Conf. on Embedded Computer Systems*, pp. 93–101 (2010).
35. M. Humenbergera et al., "A fast stereo matching algorithm suitable for embedded real-time systems," *Comput. Vis. Image Underst.* **114**(11), 1180–1202 (2010).
36. S. Hadjitheophanous et al., "Towards hardware stereoscopic 3D reconstruction a real-time FPGA computation of the disparity map," in *Design, Automation & Test in Europe Conf. & Exhibition*, pp. 1743–1748 (2010).
37. M. Buder, "Dense real-time stereo matching using memory efficient semi-global-matching variant based on FPGAs," *Proc. SPIE* **8437**, 843709 (2012).

Hejian Li received her bachelor's degree in physics from East China University of Science and Technology, Shanghai, in 2000 and her master's degree in microelectronics from Fudan University, Shanghai, in 2007. She is currently pursuing her PhD degree in the Key Laboratory of Ministry of Education for Advanced Display and System Application, Shanghai University. She had been a lead engineer in Cadence. Her research interests include three-dimensional video processing, electronic design, image/video processing, and coding.

Ping An received her BS in radio technology and MS in signal and information processing from Hefei University of Technology, Hefei, China, in 1990 and 1993, respectively, and her PhD degree in communication and information systems from Shanghai University, Shanghai, in 2002. She is currently a professor in the School of Communication & Information Engineering, Shanghai University. Her research interests include stereoscopic and three-dimensional vision analysis, image and video processing, coding, and application.

Zhaoyang Zhang received his BS degree from the Department of Wireless Electronic Engineering, Xi'an Jiaotong University, Xi'an, China, in 1962. He is currently a professor and PhD supervisor of Shanghai University. He is the vice director of the Key Laboratory of Advanced Display and System Application, and he was vice-director of the School of Communication and Information Engineering at the University of Shanghai. His research interests include image processing, video compression and multimedia communication, and digital video and streaming techniques.