Proceedings of the 2001 IEEE
International Conference on Robotics & Automation
Seoul, Korea • May 21-26, 2001

VLSI Processor for Reliable Stereo Matching Based on Adaptive Window-Size Selection

M. Hariyama, T. Takeuchi and M. Kameyama
Department of Computer and Mathematical Sciences
Graduate School of Information Sciences
Tohoku University Aoba 05, Aramaki, Aoba-ku
Sendai 980-8579, Japan

1 Introduction

Acquisition of reliable three-dimensional(3-D) images of a real scene plays an essential role in real-world intelligent systems such as intelligent robots and highly-safe vehicles. Stereo vision is a well-known method to acquire three-dimensional information. One important problem on stereo vision is to establish reliable correspondence between images. Another problem is that the correspondence search is time-consuming even if state-of-art general-purpose processors are used to accelarate the corresponding search. From this point of view, this paper presents a reliable stereo-matching algorithm and a new parallel VLSI processor architecture for stereo matching.

One commonly-used method to establish correspondence between images is the SAD(sum of absolute differences) method. Given a pixel L in one image(reference image), an SAD is computed between a rectangular window centered at L and a candidate window of each possible location in another image (candidate window). In usual cases, the window size are empirically pre-determined, and the candidate pixel with the smallest SAD is determined to be the corresponding pixel of L. The major problem on the SAD-based matching is that a window size for SAD computation must be large enough to avoid ambiguity but small enough to avoid the effects of projective distortions[1]. To solve this problem, several algorithms have been reported until now[2]-[3]. However, these algorithms are not always the best approaches in reliablitty and parallellism, since regularity and parallelism are not sufficient to implement a super-performance VLSI processor.

From this point of view, we propose a VLSI-oriented stereo matching algorithm with variable window sizes. The method is mainly based on an idea that an SAD graph has a unique and clear minimum at the reliable matching pixel. First, a window size is iteratively enlarged to select as small a window for each pixel as possible that can avoid ambiguity based on uniqueness of a minimum of an SAD graph. This process is called a global search. Next, the estimate of the corresponding pixel obtained by the global search is iteratively refined by shrinking the window size. To avoid ambiguity with a small window size, the correspondence estimate obtained by the global search is efficiently used.

The proposed algorithm has regular data flow based on iterations of SAD computation so that it is suitable for parallel processing. In designing its VLSI processor, the major consideration is to achieve high utilization of processing elements(PEs) for SAD computation. In SAD computation, a degree of parallelism between pixels in a window changes depending on its window size. Pixel-parallel SAD computation results in low utilization since many PEs may not be utilized for a small window size.

To solve this problem, an SAD is computed in a pixel-serial manner where a single absolute difference(AD) is computed in each control step. The regular data flow of the pixel-serial computation makes it possible to fully utilize a PE for SAD computation. Moreover, in a correspondence search, a degree of window-level parallelism is predetermined by an image width. Therefore, candidate windows of the equal number are assigned to each PE in advance so that PEs are fully utilized.

The processing time of the stereo matching VLSI processor designed in a 0.5μm CMOS rule is evaluated to be $60ms$ for a pair of input images of a size 512×512. The performance is more than ten thousand times higher than that of the general-purpose microprocessor(Pentium II 400MHz). The highest perfomance as well as highest reliablity of stero matching will be achieved among the processors reported until

now.

2 Stereo Matching Algorithm

2.1 Basic Stereo Matching Algorithm

Once correspondence between images is established, a 3-D point in the real scene can be found by triangulation. To establish the correspondence, a similarity measure must be computed which reflects how well a pixel L in the left image matches each pixel on the epipolar line in the right image as shown in Fig. 1. One commonly used similarity measure is a sum of absolute differences(SAD). Let us consider a reference window of a size $W \times W$ centered at $L(=(U_L, V_L))$ in the right image and a candidate window centered at (U_R, V_R) on the epipolar line in the left image as shown in Fig. 1. Then, an SAD in a window size W is given by

$$F_W = \sum_{j=-\frac{W-1}{2}}^{\frac{W-1}{2}} \sum_{i=-\frac{W-1}{2}}^{\frac{W-1}{2}} |I_L(U_L+i, V_L+j) - I_R(U_R+i, V_R+j)| \quad (1)$$

where I_L and I_R are intensity values in the left and right images, respectively. If a candidate window exactly matches the reference window, then the SAD becomes 0. Given a reference pixel L in the left image, an SAD is computed for each candidate pixel on the epipolar line in the right image, and an SAD curve is obtained as shown in Fig. 2. A pixel where the SAD curve has its minimum is called a "*matching*" pixel. In a straightforward method, a window size is empirically predetermined. The matching pixel in the window size is determined as the corresponding pixel.

The window size is an important parameter in SAD-based stereo matching. If the window size is too small, there exist several possibilities for the choice of the corresponding pixel. Therefore, the window size must be large enough to avoid the ambiguity. On the other hand, if the window size is too large and the window includes pixels whose depths in the scene are different from each other, the matching pixel may not be the corresponding pixel due to different projective distortions in the left and the right images.

2.2 Reliable Stereo Matching with Variable Window Sizes

To solve this problem, we propose stereo matching algorithm that adaptively select a window size for each pixel. The algorithm consists of two following steps.

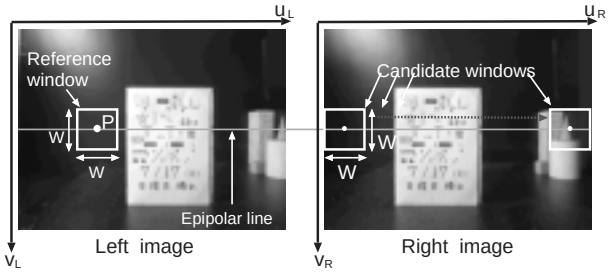

Figure 1: Search for a corresponding point.

2.2.1 Global Search Based on a Uniqueness Measure of SAD graphs

As shown in Fig. 2(a), an SAD graph usually has multiple local minima. Let Q_1^W be a matching pixel in a window size W. Let Q_2^W be a pixel where the SAD graph has the second smallest value of all the local minima. Then, a reliability measure R_W of the matching pixel is given by

$$R_W = \frac{(F_W(Q_1^W) - F_W(Q_2^W))}{W^2}.$$

The similarity measure is based on an idea that the matching pixel Q_1^W is reliable if the SAD graph has a unique and clear minimum at Q_1^W.

Based on the similarity measure, the window size is iteratively enlarged to select as small a window as possible that will still produce the reliable matching pixel. In the global search, following steps are repeated until the corresponding pixel are determined, or the window size becomes larger than the maximum window size.

Step 1. Compute the matching pixel Q_1^W and the reliability measure R_W from an SAD graph in a window size W(Fig. 2(a)), where W is initially set to the minimum window size.

Step 2. Check whether the matching pixel Q_1^W is reliable or not. From an SAD graph in a window size $W+2$, the matching pixel Q_1^{W+2} and the reliability measure R_{W+2} are computed(Fig. 2(b)). Only if following two conditions are satisfied, then the matching pixel Q_1^W is reliable and determined as the corresponding pixel.

- The pixel Q_1^W is the matching pixel in both window sizes W and $W+2$.
- The reliability measure increases with the window size.

Otherwise, W is set to $W+2$.

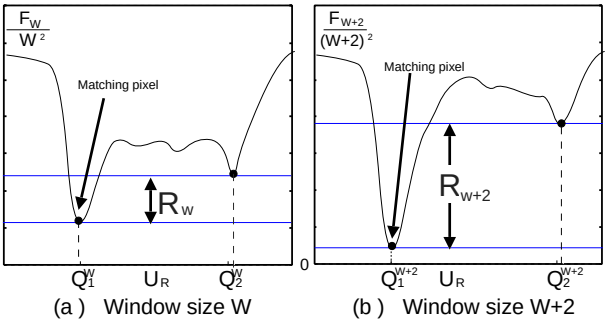

Figure 2: SAD graphs for window sizes W and $W+2$.

2.2.2 Local Search

Since the window size is enlarged in the global search, the result of the global search would include errors due to differences between projective distortions in the left and the right images. To overcome this problem, the result of the global search is iteratively refined by shrinking the window size. The major issue of the local search is to shrink the window size without ambiguity. For the purpose, we introduce the smoothness constraint. For example, let us consider the local search for the corresponding pixel of a pixel P in the left image as shown in Fig. 3. Let be W be the window size determined by the global search. Let $D_{i,j}$ $(1 \leq i, j \leq W)$ be the disparities for the pixels in the reference window. The smoothness constraint is based on the idea that the world is mostly made up of objects with smooth surfaces. If the 3-D pixel projected onto P is on a smooth surface in the real scene, the probability for the correct disparity of P is included in $D_{i,j}(1 \leq i, j \leq W)$ becomes large. Based on this observation, correspondence search with the window size $W/2$ is done for only candidate pixels with the disparity $D_{i,j}(1 \leq i, j \leq W)$, and the candidate pixel with the minimum SAD is selected as a corresponding pixel in the window size. Although there exists several candidate pixels with the same SADs, ambiguity can be resolved by this limitation of the candidate windows. The above steps to shrink the window size are repeated until the window size reaches the minimum window size.

2.2.3 Evaluation

Figure 4 (a) and (b) show 3-D plots of a front-parallel rectangular plane(Fig. 1) obtained by the fixed-window-size method and the variable-window-size method, respectively. Figure 4 (c) shows a comparison result of the depth error. The result shows

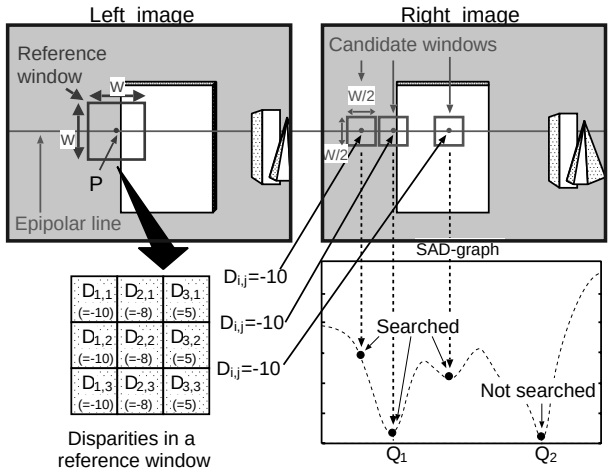

Figure 3: Local search.

that the variable-window-size method selects reliable corresponding points.

Another advantage of our method is that it does not require any empirically-determined parameter. Therefore, it is suitable not only for static environment but also for mobile applications in dynamically changing environment.

3 Pixel-Serial and Window- Parallel Architecture

3.1 Overview

Figure 5 shows a block diagram of the stereo vision VLSI processor. It mainly consists of two Laplacian-of-Gaussian(LOG) filter units, two image memories, buffers for a reference window and candidate windows, and a SAD unit. To increase reliability of the stereo matching, Laplacian-Of-Gaussian(LOG)-filtered images are used as input images for SAD computation instead of original images. LOG filtering is relatively time-consuming so that it is implemented by using field-programmable gate arrays(FPGAs). FPGAs are useful to implement special-purpose processor cost-effectively since their functions can be changed by programming. A gray rectangle denotes the units integrated on the full-custom VLSI chip. A capacity of each image memory is too large to integrate them and the stereo matching unit on a single chip. In a typical case, each image memory has 256K-byte capacity for a 256-level gray-scale image with a size of 512×512. Therefore, images are stored in external memories. The external memories cause a data-

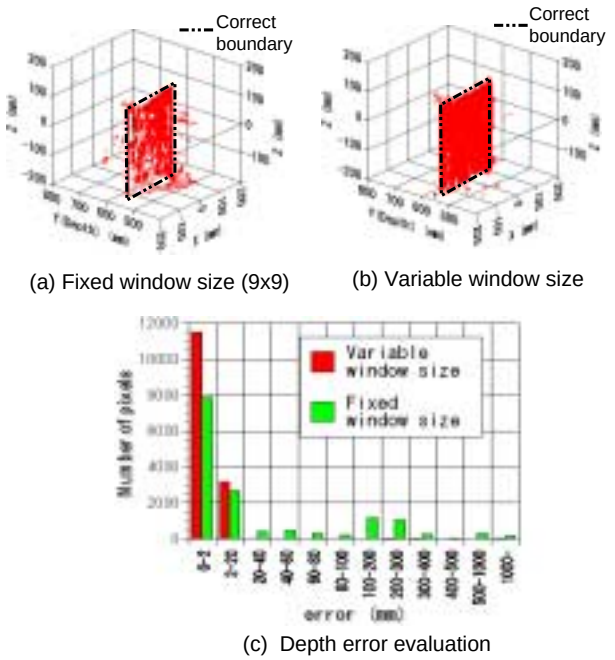

Figure 4: Evaluation for the front-parallel rectangular plane.

transfer bottleneck due to its large access time. To solve this problem, frequently used pixels are stored in on-chip buffers with smaller access time.

A corresponding search is performed as follows. Firstly, a reference window and candidate windows on an epipolar line are retrieved from image memories, and they are stored in buffers. Secondly, a corresponding pixel is searched in the SAD unit. Finally, the resulting two-dimensional(2-D) coordinates of a corresponding pixel is send to a host processor that computes 3-D coordinates from the 2-D coordinates based on triangulation. The 3-D coordinate computation is executed by the host processor since its computational amount is small. The above mentioned steps are repeated for all the reference windows. Moreover, all the steps are overlapped in execution by pipelining to obtain the highest performance.

3.2 Stereo Matching Unit Based on a Pixel-Serial and Window-Parallel Scheduling

In the search for the corresponding pixel, there exist pixel-level parallelism and window-level parallelism as follows:

Window-level parallelism. SADs can be computed in parallel for all the candidate windows on the epipolar line as shown in Fig. 6.

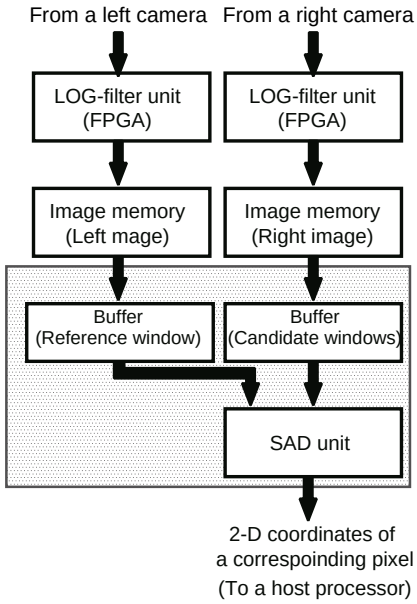

Figure 5: Overview of a stereo vision VLSI processor.

Pixel-level parallelism. Absolute differences (ADs) in Eq. (1) can be computed in parallel for all the pixels in a candidate window.

To achieve the highest-performance under a hardware-resource constraint, these parallelism should be carefully combined or selected.

For window-size-fixed stereo matching, pixel-level parallelism can be fully exploited. Figure 7 shows a pixel-parallel and window-serial (PPWS) scheduling that is usually used[4]. However, pixel-level parallelism is not suitable for the window-size-variable stereo matching due to low utilization of hardware. Figure 8 shows its simple example. Assumed that there exists hardware for pixel-parallel SAD computation for 2×2 window as shown in Fig. 8(a). When a window size is 1×2, only two AD circuits and one adder are utilized. In a typical case, the maximum window size and the minimum window size are 25 and 3, respectively. Hardware for computing $625(25 \times 25)$ ADs are required to exploit the pixel-level parallelism fully. However, only hardware for computing $9(= 3 \times 3)$ ADs are utilized when computing an SAD for a 3×3 candidate window.

To solve this problem, the pixel-serial and window-parallel (PSWP) scheduling is proposed, where an SAD for a candidate window is computed in a pixel-serial manner, and SADs for different candidate windows are computed in parallel as shown in Fig. 9.

Figure 10 shows a block diagram of a processing el-

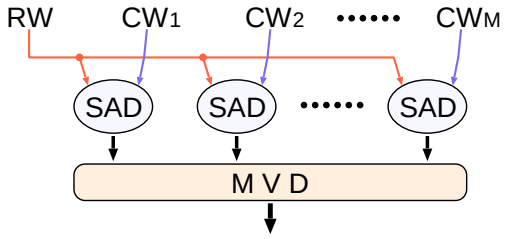

Figure 6: Data-flow graph of correspondence matching.

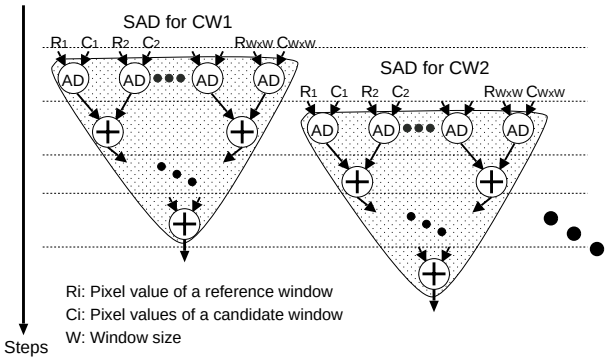

Figure 7: Data-flow graph of the PPWS SAD computation.

ement(PE) that computes an SAD in the pixel-serial manner. An AD circuit and an adder in the PE can be 100% utilized since an AD and an addition are computed in each step in the pixel-serial scheduling independently of the window size W. Figure 11 shows an overall architecture of a stereo matching unit based on the PSWP scheduling. It consists of M PEs to computes SADs for M candidate windows in parallel. Pixels in the right image are stored in a memory module ML. For parallel access, pixels in the right image are equally distributed among memory modules $MR_i(i = 1, \cdots, n)$. Since the number of candidate windows on the epipolar line is fixed in advance, SAD computation for candidate windows of the same number can be assigned to each PE in advance so that all the PEs are 100% utilized. The minimum-value-determination unit computes the reliability measure R from the computed SADs and determines a corresponding pixel.

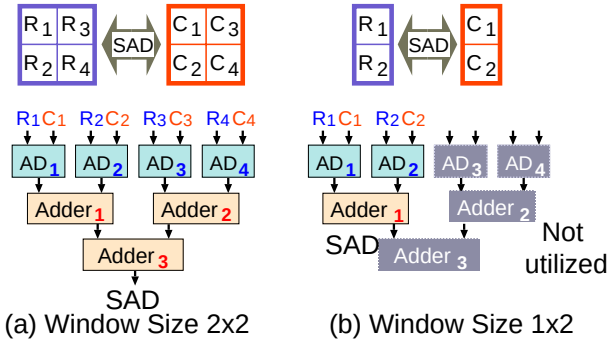

Figure 8: Problem on the PPWS scheduling.

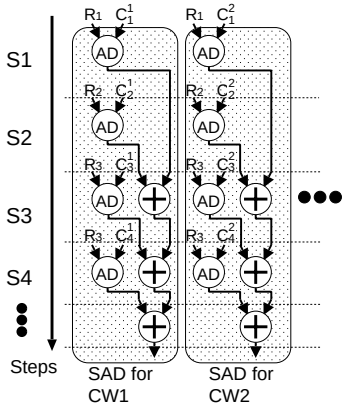

Figure 9: Data-flow graph of the PSWP SAD computation.

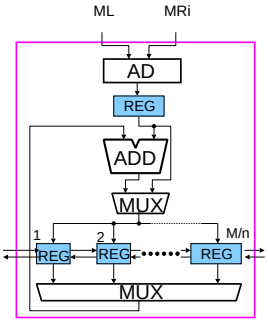

Figure 10: PE for the pixel-serial SAD computation.

3.3 Evaluation

Figure 12 shows a floor plan and a chip layout of the stereo vision VLSI processor designed in a 0.5μm CMOS process. It consists of an SAD unit, a candidate buffer, a reference buffer, a minimum value detection unit and a control unit. The candidate buffer has 512×26-byte capacity since 256-level LOG-filtered images are used as left and right images, and the maximum window size W_{max} is set to 25. Features of the VLSI processor are summarized in Table 1. The time required to produce a depth map estimated to be $60m$sec for input images of a size 512×512. The performance of the VLSI processor is more than ten thousand times faster than the general-purpose processor(Pentium II 400MHz). The performance can be increased in proportion to the number of the VLSI processors since correspondence search for different reference windows can be performed in parallel.

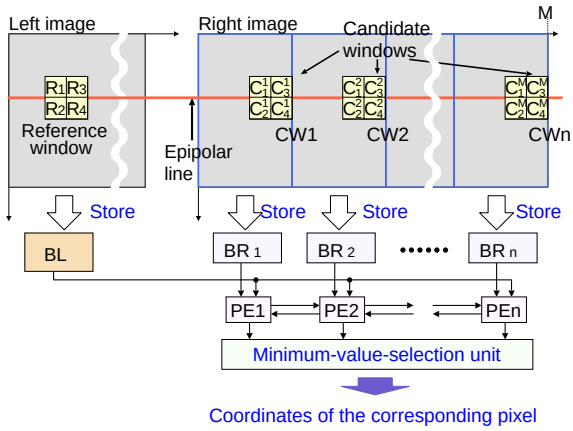

Figure 11: Block diagram of the stereo matching based on the PSWP scheduling.

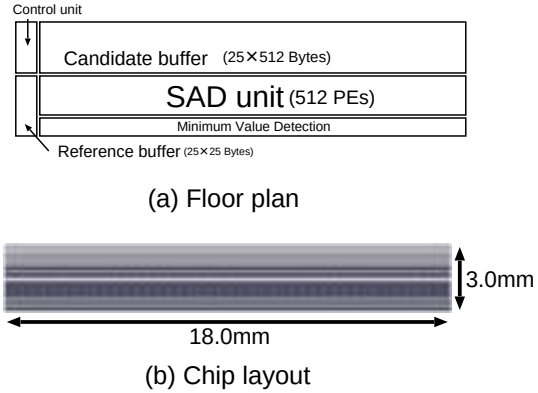

Figure 12: Layout of the stereo matching unit.

4 Conclusion

The variable-window-size method determines the reliable matching pixel using only SAD graphs. Therefore, to increase reliability of stereo matching, it can be easily combined with the multi-baseline stereo that uses a sum of SADs as a similarity measure[5].

The window-parallel and pixel-serial architecture allows a simple interconnection network between PEs and memory modules. This nature is suitable not only for the full-custom VLSI processor but also the FPGA-based processor where the performance is seriously decreased by the interconnection delay. Figure 13 shows a prototype of the FPGA version using Xilinx Virtex1000 FPGAs and an Aptix MP3C system. Its performance becomes two thousand times higher than the general-purpose microprocessor.

Table 1: Features of the VLSI processor

Technology	0.5-μm CMOS double-metal process
Area	$18.0 \times 16.8\text{mm}^2$
Input image	512 × 512 pixels(256-level gray scale)
Maximum window size	25
Performance	60 $m$sec/depth map
Number of transistors	1 300 000
Clock frequency	200MHz
Supply voltage	5V

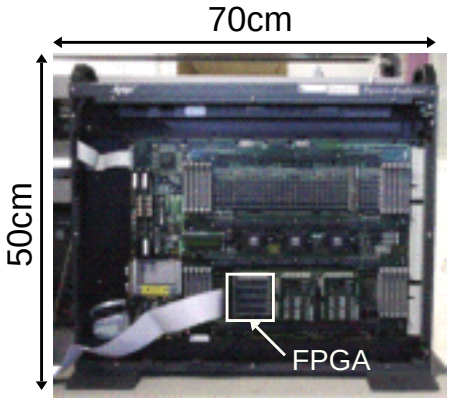

Figure 13: FPGA version of the stereo vision VLSI processor.

References

[1] S.T.Barnard and M.A.Fischler, "Stereo vision, " in Encyclopedia of Artificial Intelligence. New York: John Wiley, pp.1083-1090, 1987.

[2] T. Kanade and M. Okutomi, "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment," IEEE Trans. PAMI, vol.16, no. 9, pp.920-932, 1989.

[3] D. Scharstein and Richard Szeliski, "Stereo Matching with Non-Linear Diffusion," in Proc. CVPR, pp.343-350, 1996.

[4] S. Lee , M. Hariyama, M. Kameyama, "A Three-Dimensional Instrumentation VLSI Processor Based on a Concurrent Memory-Access Scheme," IEICE Trans. Electron, vol.E80-C, No.11, pp.1491-1498,1997.

[5] M. Okutomi, T. Kanade, "A Multiple-Baseline Stereo," IEEE Trans. PAMI, vol.15, no.4, 1993.