Cost Aggregation and Occlusion Handling With WLS in Stereo Matching

Dongbo Min, *Student Member, IEEE*, and Kwanghoon Sohn, *Member, IEEE*

***Abstract*—This paper presents a novel method for cost aggregation and occlusion handling for stereo matching. In order to estimate optimal cost, given a per-pixel difference image as observed data, we define an energy function and solve the minimization problem by solving the iterative equation with the numerical method. We improve performance and increase the convergence rate by using several acceleration techniques such as the Gauss–Seidel method, the multiscale approach, and adaptive interpolation. The proposed method is computationally efficient since it does not use color segmentation or any global optimization techniques. For occlusion handling, which has not been performed effectively by any conventional cost aggregation approaches, we combine the occlusion problem with the proposed minimization scheme. Asymmetric information is used so that few additional computational loads are necessary. Experimental results show that performance is comparable to that of many state-of-the-art methods. The proposed method is in fact the most successful among all cost aggregation methods based on standard stereo test beds.**

***Index Terms*—Cost aggregation, multiscale approach, occlusion handling, stereo vision, weighted least square.**

I. Introduction

FOR decades, the correspondence problem has been an important issue in the field of computer vision, and many methods have been proposed to solve this problem. Given two or more images of the same scene taken from different viewpoints, the correspondence problem is to find the corresponding points in other images for a pixel in one image and compute the disparity map which is a set of the displacement vectors between the corresponding pixels. In binocular stereo, it is assumed that two input images are calibrated and rectified in order to make the problem easy and accurate, so that epipolar line becomes horizontal and the search range is limited in a horizontal direction. Dense disparity maps acquired by stereo matching can be used in many applications, including image-based rendering, 3-D object modeling, robot vision, tracking, etc. An extensive review of stereo matching algorithms can be found in [1]. Stereo matching presents a difficult problem due to image ambiguity. This ambiguity may include noise factors, occlusion, and lack of texture. In order to reduce the ambiguities and uncertainties of matching, many algorithms have been proposed using several constraints. Generally, stereo matching algorithms can be classified into two approaches (global and local) based on the strategies used for estimation. Global approaches define energy models that apply various constraints to reduce uncertainties of disparity maps and solve them through various minimization techniques (such as graph cut and belief propagation). Local approaches use correlations between color or intensity patterns in neighboring windows. These approaches can easily obtain correct disparities in highly textured regions. However, they often tend to produce noisy results in large untextured regions. Moreover, they assume that all the pixels in a matching window have similar disparities, resulting in blurred object boundaries and the removal of small details. Performance depends on how the optimal window is selected in each pixel, but finding an optimal window with an arbitrary shape and size is very difficult. This is generally known as the NP-hard problem. To define the optimal window for each pixel well, it is necessary to be aware of depth information, which must be estimated. To solve this problem, a number of methods have been proposed.

In general, adaptive window algorithms try to find optimal windows for each pixel by adaptively changing the window size and shape. Kanade and Okutomi [2] proposed a way of selecting an appropriate window by evaluating the local intensity and disparity variations. This method is, however, highly dependent on the initial disparity estimation and is computationally expensive. Moreover, the window shape has to be constrained as a rectangle due to the high computational complexity. Boykov [3] proposed a way of performing plausibility testing and computing accurate windows for each pixel. Veksler [4] presented an algorithm that could be used to select a certain window size and shape over a large class of many windows. Efficient optimization over many windows was achieved using the minimum ratio cycle algorithm for graphs. However, the window shape is not general and the computation of window cost requires many parameters. For further efficient implementation, a fast method was proposed with integral images and dynamic programming [5]. Segmentation-based cooperation works by selecting the size and shape of windows by using the segmentation results for reference images [6]–[8]. These approaches yield good results, but they require color segmentation, which may cause high computational complexity. Gong [9] proposed the adaptive cost aggregation scheme using edge detection instead of color segmentation for real-time implementation with graphic hardware (GPU). Tang [10] presented a method of region growing with an adaptive window, but the weight of the window depended on the geometric distance only. Yoon *et al.* [11] used boundary information to compute accurate windows for each pixel, and solved

Manuscript received April 30, 2007; revised April 13, 2008. First published June 17, 2008; last published July 11, 2008 (projected). This work was supported in part by MEST, MKE, and MOLAB through the fostering project of the Lab of Excellency, and in part by the MKE, Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA [IITA-2008-(C1090-0801-0011)]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hassan Foroosh.

The authors are with the School of Electrical and Electronic Engineering, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul, 120–749, Korea (e-mail: forevertin@yonsei.ac.kr; khsohn@yonsei.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2008.925372

1057-7149/$25.00 © 2008 IEEE

the iterative problem in a hierarchical framework, known as the scale-variant iterative scheme.

Multiple window algorithms evaluate a small number of different windows, whose reference points lie in several positions, and then select the one with the lowest cost [12]. Okutomi *et al.* [13] used multiple stereo pairs and windowing to solve the boundary overreach and occlusion problems.

Yoon and Kweon [14] proposed a general method that computes an optimal local support window. It uses a fixed-sized support window with varying weights for each pixel. Although no global optimization techniques were used in the original research, it yielded very strong results. However, in general, it is very computationally expensive to perform pixel-wise support weight computation. Wang *et al.* [15] introduced the adaptive weight approach in a dynamic programming framework, and applied the method into a real-time approach, by using GPU hardware. The matching cost is aggregated over the vertical 1-D window only. They also proposed a two-pass approximation scheme to accelerate the original method.

Another issue discussed in this paper is occlusion handling. For stereo image pairs, occluded pixels are usually only visible in one image so it is impossible to estimate disparity information from the occluded pixels. However, in many applications such as image-based rendering and 3-D modeling, it is necessary to assign reasonable disparity fields to the occluded pixels. Several constraints have been used in stereo matching for occlusion handling. The ordering constraint preserves the order of matching along the scanline in stereo image pairs. Approaches which exploit the ordering constraint have used dynamic programming [16]. Stereo matching has been formulated as the problem of finding a minimum cost path in the matrix of matching costs between the corresponding scanlines. Dynamic programming finds the minimum path for each scanline. In the estimated minimum path, the discontinuities correspond to the left and right occlusions. This approach is very efficient but the ordering constraint is invalid when an image has a thin object. Moreover, to generate the cost matrix it is necessary to define an appropriate penalty for occluded pixels, which may be sensitive to the matching cost. The uniqueness constraint means that the corresponding points between the two images are unique. According to this principle, each pixel must have at most one disparity vector. A simple method of detecting occlusion with the uniqueness constraint is the cross-checking technique. Many approaches have been proposed to estimate the disparity of the occluded pixels by combining the uniqueness constraint into a global optimization method [17], [18]. Most approaches have used iterative schemes to detect the occlusion regions and assign predefined values to the occluded pixels, which may also be sensitive to the matching cost. They estimate the symmetric disparity fields for the left and right images, which may cause the computational complexity to double.

In this paper, we propose a novel approach of performing efficient cost aggregation and handling occlusion for stereo matching. Most conventional cost aggregation methods have tried to find the optimal window by using different, predefined windows or by computing the weight of the window based on disparity and intensity information. To estimate the optimal cost, given a per-pixel difference image as observed data, we define an energy function and solve the minimization problem by solving a corresponding iterative equation with the numerical method. In this case, since the energy function is convex, its minimization problem does not suffer from being trapped in local minima. For occlusion handling, it is not necessary to define a predefined value for the occluded pixel and it is possible to use asymmetric information, that is, only the left disparity field needs to be used. Therefore, it is possible to estimate the disparity fields for the occluded pixels efficiently and correctly. The main contributions of the proposed method are as follows.

1) We define an energy function for optimal cost aggregation and solve the minimization problem using the numerical method.
2) When solving the iteration scheme, we use several acceleration techniques to improve performance and increase the convergence rate.
3) We combine the occlusion problem, which has not been solved by any existing cost aggregation approaches, into the iterative scheme. We only use asymmetric information for occlusion handling with trivial additional computational loads.

The minimization problem can be converted into nonlinear diffusion filtering while respecting the discontinuities. Nonlinear filtering is used to solve various problems including image restoration/denoising, motion estimation, stereo matching, etc. In stereo matching, many approaches have been proposed to estimate the disparity field which has good discontinuity localization and is robust to textureless regions [19], [20]. The proposed method differs from conventional methods in the sense that nonlinear filtering is executed in the matching cost domain, not in the image domain. Therefore, the minimization problem is not trapped in local minima, which is one of the most serious problems of stereo matching or motion estimation when using the variational method. We manage to solve the occlusion problem by combining it into the iterative scheme to minimize the proposed energy function.

The remainder of this paper is organized as follows. In Section II, we discuss the proposed method, and then introduce the cost aggregation scheme with anisotropic diffusion and weighted least square and the acceleration methods in Section III. The occlusion handling methods are described in Section IV. Finally, we present the experimental results and conclusion in Sections V and VI, respectively.

II. Problem Statement

When estimating the disparity field, only the left and right image pairs are used. We obtain the difference image by shifting the right image further to the right, and then subtracting the left and the shifted right images. This is done for all disparities. A set of difference images is called 3-D cost volume $e(p, d)$, where p and d represent the 2-D locations of pixels and disparity, respectively. We call the 3-D cost volume a per-pixel cost. In order to estimate the optimal cost, we define the per-pixel cost e as follows:

$$e(p, d) = E(p, d) + n \tag{1}$$

where n represents noise. We simplify $E(p, d)$ to $E(p)$, since the same process is performed for each disparity. In other

(a) (b) (c)

Fig. 1. Per-pixel and estimated costs in the "Tsukuba" image: (a) stereo image pairs, (b) per-pixel and estimated cost when disparity is 0, (c) per-pixel and estimated cost when disparity is 6. The estimated cost is computed with the proposed method, and λ is set to 1.0. We can find that the per-pixel cost is very noisy.

words, $E(p)$ can be referred to as the 2-D cost function which is the section of 3-D cost volume, given the disparity value of d. Fig. 1 shows the per-pixel and estimated costs for the "Tsukuba" image. We find that the per-pixel cost function is very noisy. Given the observation data, which represents a per-pixel cost, we use the prior knowledge that costs should vary smoothly, except at object boundaries. From this observation, we are able to estimate the cost function by minimizing the following energy model:

$$\varepsilon(E) = \int_{\Omega} (E(p) - e(p))^2 dp + \lambda_d \int_{\Omega} g(|\nabla I|)|\nabla E|^2 dp \quad (2)$$

where $g(|\nabla I|)$ decreases monotonically with respect to $|\nabla I|$. This is known as the diffusivity function, which plays the role of a discontinuity marker. λ_d is the weighting factor used to control the ratio between the fidelity and smoothness terms.

The proposed scheme is different from the nonlinear diffusion scheme for cost aggregation, proposed by Scharstein *et al.* [21] and derived from a Bayesian model of stereo matching. While this scheme performs nonlinear diffusion based on the MRF model by using the cost function with different disparities (diffusion over disparity), the proposed method performs diffusion by using the intensity of an image in the cost function with the same disparity (diffusion in disparity).

III. Proposed Cost Aggregation

A. Cost Aggregation With Anisotropic Diffusion

Anisotropic diffusion filtering is one of the most popular approaches in image restoration/denoising, image segmentation, 3-D reconstruction, motion estimation, stereo matching, etc. Especially in stereo and motion analysis, the minimization problem is solved with scale-space filtering to avoid becoming trapped in local minima, since the minimization problem is nonconvex. The proposed method contains a convex energy function since the data term is quadratic, as shown in (2). The minimization of (2) yields the following Euler–Lagrange equation:

$$E(p) - e(p) - \lambda_d \nabla \cdot (g(|\nabla I|)\nabla E) = 0. \quad (3)$$

We obtain the solution to the equation by calculating the asymptotic state $(t \to \infty)$ of the parabolic system, as shown in (4)

$$\frac{\partial E}{\partial t} = -E(p) + e(p) + \lambda_d \nabla \cdot (g(|\nabla I|)\nabla E). \quad (4)$$

(a) (b)

Fig. 2. Results for anisotropic diffusion in the "Venus" image: (a) per-pixel and estimated cost when disparity is 0, (b) estimated disparity map obtained by WTA.

We then discretize the equation to find the solution using a finite difference method. All spatial and temporal derivatives are approximated by forward differences and a semi-implicit scheme. The final iterative scheme is as follows:

$$E^{t+1} = \frac{1}{1+\mu} E^t + \frac{\mu}{1+\mu} e + \frac{\mu \lambda_d}{1+\mu} \nabla \cdot (g(|\nabla I|)\nabla E^t). \quad (5)$$

We use the diffusivity function proposed by Geman and Mcclure [26]. In (5), we define the step size as $\mu = 0.1$ and the weighting factor as $\lambda_d = 10$. Fig. 2 shows the results of the estimated cost with nonlinear diffusion filtering, and the disparity field computed with these cost functions. To evaluate the performance of the cost aggregation scheme, we use the winner-takes-all (WTA) method as an optimization technique. We find that the estimated disparity field is not satisfactory. Although edge-preserving smoothing is performed and the impulsive noise generated by the sensor is reduced, there are some problems in the textureless and occluded regions. In (1), the optimal cost model considers sensor noise only. Thus, another strategy to solve these problems is necessary.

B. Cost Aggregation With Weighted Least Square

To solve the problem in textureless regions, we consider using the smoothness constraint with more neighborhoods. As opposed to image restoration/denoising, it is necessary to gather sufficient texture in the neighborhoods for reliable matching. To include more neighborhoods, we propose a new energy function with the weighted least square method

$$\varepsilon(E) = \int_{\Omega} (E(p) - e(p))^2 dp + \lambda \int_{\Omega} \sum_{n \in N_1} \left\{ \begin{array}{c} w(p, p+n)(E(p) - E(p+n))^2 \\ +w(p, p+n^{\perp})(E(p) - E(p+n^{\perp}))^2 \end{array} \right\} dp$$
$$N_1 = \{(x_n, y_n) | 0 < x_n \leq M, 0 \leq y_n \leq M\} \quad (6)$$

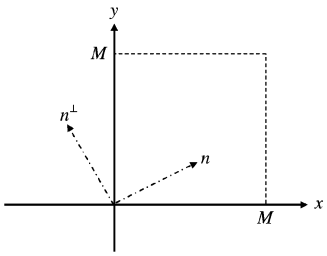


Fig. 3. Neighboring pixels used in the smoothness term: when n contains (1,0) only, it is similar to anisotropic diffusion.

where w represents the weighting function between corresponding neighbor pixels, and is defined with color or intensity information. n and $n^{\perp}$ represent the 2-D vectors, which are perpendicular to each other. M represents the size of a set of neighbor pixels. When the element of the set N_1 is (1,0) only, (6) is similar to anisotropic diffusion. In other words, (6) can be considered to be the generalized function of (2). Fig. 3 shows the neighboring pixels which are used in the weighted least square method. Taking the first derivative of (6) with respect to E, we yield the following equation:

$$\begin{aligned} &E(p) - e(p) \\ &\quad + \lambda \sum_{n \in N_1} \left\{ \begin{array}{c} \nabla_{-n} w(p, p+n)(E(p) - E(p+n)) \\ +\nabla_{-n^{\perp}} w(p, p+n^{\perp})(E(p) - E(p+n^{\perp})) \end{array} \right\} \\ &\quad = 0 \end{aligned} \tag{7}$$

where ∇_{-n} and $\nabla_{-n^{\perp}}$ represent the first derivative operators along $-n$ and $-n^{\perp}$, respectively. We induce the above equation as follows:

$$\begin{aligned} &E(p) - e(p) \\ &\quad + \lambda \sum_{n \in N_1} \left\{ \begin{array}{c} w(p, p+n)(E(p) - E(p+n)) \\ -w(p-n, p)(E(p-n) - E(p)) \\ +w(p, p+n^{\perp})(E(p) - E(p+n^{\perp})) \\ -w(p-n^{\perp}, p)(E(p-n^{\perp}) - E(p)) \end{array} \right\} = 0. \end{aligned} \tag{8}$$

To simplify the above equation, we redefine the set of neighbor pixels. When p is (x, y), the set can be expressed as

$$N(p) = \{(x+x_n, y+y_n) | -M \le x_n, y_n \le M, x_n + y_n \ne 0\}.$$

By using the above notation, (8) is expressed as

$$E(p) - e(p) + \lambda \sum_{m \in N(p)} w(p, m)(E(p) - E(m)) = 0. \tag{9}$$

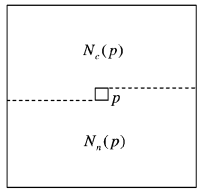


Fig. 4. Causal and noncausal parts used in Gauss–Seidel acceleration.

In the above equation, we assume that the weighting function w is symmetric

$$w(p, q) = w(q, p). \tag{10}$$

The solution of the $(k+1)$th iteration is obtained by the following equation:

$$\begin{aligned} E^{k+1}(p) &= \bar{e}(p) + \bar{E}^{k}(p) \\ &= \frac{e(p) + \lambda \sum_{m \in N(p)} w(p, m) E^{k}(m)}{1 + \lambda \sum_{m \in N(p)} w(p, m)}. \end{aligned} \tag{11}$$

Equation (11) consists of two parts: normalized per-pixel matching cost and weighted neighboring pixel cost. By running the iteration scheme, the cost function E is regularized with the weighted neighboring pixel cost. The iteration scheme is similar to the adaptive weight approach when the number of iterations is 1. In the proposed method, we use the symmetric Gaussian weighting function with the CIE-Lab color space in (12). r_c and r_s are weighting constants for the color and geometric distances, respectively. As opposed to $g(|\nabla I|)$ in (3) it is necessary to use the term for geometric distance in the weighting function, since the smoothness constraints with more neighborhoods are considered

$$\begin{aligned} &w(p, m) \\ &\quad = \exp\left(-\left(\frac{C_L(p, m)}{2r_c^2} + \frac{C_R(p, m)}{2r_c^2} + \frac{S(p, m)}{2r_s^2} \right) \right) \\ &C(p, m) = (L(p) - L(m))^2 + (a(p) \\ &\quad - a(m))^2 + (b(p) - b(m))^2 \\ &S(p, m) = (p - m)^2. \end{aligned} \tag{12}$$

C. Acceleration Scheme

1) Gauss–Seidel Acceleration: The proposed iterative scheme is accelerated with several numerical methods. One reason for slowing down the convergence in (11) is that it does not use the latest information available. The updated components in each pixel are used only after one iteration is complete. Therefore, we compensate for this problem by using

$$\begin{aligned} E^{k+1}(p) &= \bar{e}(p) + \bar{E}^{k}(p) \\ &= \frac{e(p) + \lambda \sum_{m \in N_c(p)} w(p, m) E^{k+1}(m) + \lambda \sum_{m \in N_n(p)} w(p, m) E^{k}(m)}{1 + \lambda \sum_{m \in N(p)} w(p, m)} \end{aligned} \tag{13}$$

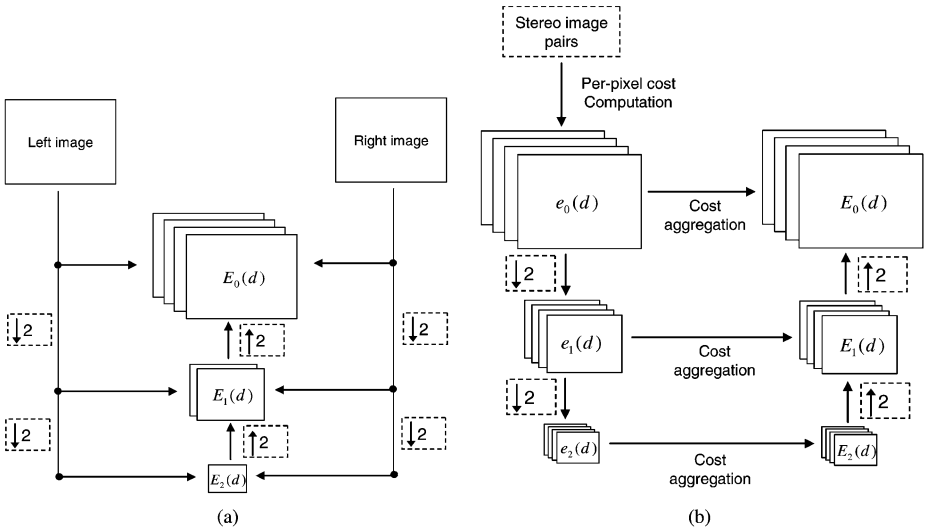


Fig. 5. Multiscale approach. (a) Conventional approach; the search range decreases when the resolution decreases to half. (b) Proposed approach.

the updated components in each pixel intermediately after they are computed. While it is generally assumed that the iterative scheme is performed from the upper left to the lower right parts of an image, we divide a set of neighbor pixels $N(p)$ into two parts as shown in Fig. 4: the causal part $N_c(p)$ and the noncausal part $N_n(p)$. Equation (11) is expressed as (13), shown at the bottom of the previous page.

Another advantage of the Gauss–Seidel acceleration method in general is that duplicate storage is not necessary for $E(p)$. Thus, we are able to make the iterative scheme more efficient [22].

2) Multiscale Approach: As previously mentioned, it is necessary to gather pixel information at a large distance to ensure reliable matching. This implies that a number of iterations are required to estimate the correct cost function. We use a multiscale approach to solve this problem. The multiscale approach has been widely used in various applications such as image segmentation, stereo matching, motion estimation, etc [23]. In stereo and motion analysis, most methods have used multiscale approaches to reduce the search range and propagate the reliable disparity and motion information [24], [25]. While conventional approaches are performed in an image domain, our method is different in the sense that the multiscale approach is applied in the cost domain to reduce the number of iterations required for the propagation of information. In (13), the cost function $E(p)$ can generally be initialized to $e(p)$. In the multiscale approach, we can initialize the value close to the optimal cost in each level by using the final value in the coarser level as the initial value in the finer level. This makes convergence faster.

Using (13), the proposed method performs cost aggregation independently in each section of the 3-D cost volume. The cost with the same disparity is used in the aggregation scheme. Conventional multiscale approaches reduce image resolution at first, and then the estimation process continues. The reduction of the resolution also reduces the search range of the disparity. For instance, if we use the multiscale approach over three levels, the search range will have been reduced to a quarter of the original search range on the coarsest level. Since the iterative scheme is performed in the cost domain, two cost functions in the finer level $E_f(p, 2d)$ and $E_f(p, 2d+1)$ are initialized by using the cost function in the coarser level $E_c(p, d)$. To avoid this problem, we use an alternative multiscale scheme for cost aggregation, as shown in Fig. 5. As this differed from conventional methods, we first compute the 3-D cost volume and then perform the proposed multiscale scheme in each 2-D cost function in a coarse-to-fine manner. Subsampling is performed on both the cost function and the stereo image pairs, since the subsampled image pairs are necessary to compute the weighting function W in each level. In order to reduce the artifacts generated during the subsampling step, we apply Gaussian lowpass filtering. The variance of the Gaussian function is proportional to the ratio of subsampling. The proposed multiscale method runs the iterative scheme at the coarsest level by initializing the cost function to $e(p, d)$. After K iterations, the resulting cost function is used to initialize the cost function in the finer level, and this process is repeated until the finest level is reached. When the cost function on the $(l+1)$th level is defined as $E_{l+1}(p)$, we compute the cost function $E_l(p)$ on the finer level by using bilinear interpolation

$$\begin{aligned} E_{l,2x,2y} &= E_{l+1,x,y} \\ E_{l,2x+1,2y} &= (E_{l+1,x,y} + E_{l+1,x+1,y})/2 \\ E_{l,2x,2y+1} &= (E_{l+1,x,y} + E_{l+1,x,y+1})/2 \\ E_{l,2x+1,2y+1} &= (E_{l+1,x,y} + E_{l+1,x+1,y} \\ &\quad + E_{l+1,x,y+1} + E_{l+1,x+1,y+1})/4. \end{aligned} \quad (14)$$

3) Adaptive Interpolation: In the previous section, we refine the resolution of the cost function by bilinear interpolation. If the cost function on the coarser level is linearly interpolated to initialize the cost in the finer level, the error can be propagated

into the neighborhood regions, especially on the boundary region. To avoid this problem, we propose an adaptive interpolation method based on the weighted least square, as follows:

$$\varepsilon(E) = \int_{\Omega} (E_l(p) - e_l(p))^2 dp + \lambda_a \int_{\Omega} \sum_{p_m \in N(p_i)} w(p, p_m)(E_l(p) - E_{l+1}(p_m))^2 dp \tag{15}$$

where $p_i = (x_i, y_i)$ represents a pixel on the coarser level, and $N(p_i)$ is defined as follows:

$$N(p_i) = \{(x_i, y_i), (x_i + 1, y_i), (x_i, y_i + 1) \quad (x_i + 1, y_i + 1) \text{ on the } l + 1 \text{ level}\}.$$

By minimizing (15), we obtain

$$E_l(p) = \frac{e_l(p) + \lambda_a \sum_{p_m \in N(p_i)} w(p, p_m) E_{l+1}(p_m)}{1 + \lambda_a \sum_{p_m \in N(p_i)} w(p, p_m)}. \tag{16}$$

In (16), w represents the weighting function, equivalent to that in (11). The even pixels are computed by using the following equation at first.

$$E_l(p) = \frac{e_l(p) + 4\lambda_a E_{l+1}(p/2)}{1 + 4\lambda_a}. \tag{17}$$

Then, the odd pixels are computed by (16). The set of neighboring pixels $N(p_i)$ consists of $(x+1, y+1)$, $(x+1, y-1)$, $(x-1, y+1)$, and $(x-1, y-1)$. Finally, the remaining pixels are computed by (16). We set the weighting factor to $\lambda_a = 15$.

Another advantage of adaptive interpolation is to increase the resolution of the cost function so that no blocking artifact exists. The adaptive interpolation by the intensity values on two successive levels leads to the up-sampling scheme, which preserves the discontinuities on the boundary region. The proposed adaptive interpolation is different from conventional image interpolation in the sense that we perform interpolation for the cost function with the intensity values that are already on the two successive levels. Thus, it is not necessary to perform the cost aggregation scheme on the finest level, and this makes the proposed method faster. In the experimental results, we will show that adaptive interpolation increases the resolution of the cost function without requiring any blocking artifact.

IV. Occlusion Handling

Most approaches have used an iterative scheme which combines the uniqueness constraint into a global optimization method for occlusion handling. This scheme assigns a predefined value to the occluded pixels, which may also be sensitive to the matching cost value. These approaches estimate both the left and right disparity fields and, therefore, double the computation complexity. In this section, we introduce a new approach for dealing with the occlusion problem in the proposed cost aggregation scheme. Only the left disparity field, and not a predefined value for the occluded pixels, is used in the occlusion handling. The occluded regions are handled by using asymmetric occlusion detection in the proposed cost aggregation method, while other methods use the symmetric matching scheme or do not handle the occluded regions at all.

Using geometric constraint only:

For each j, when $\#(S_r(j)) > 1$

Set $i_m = \arg\max_i d(S_r(j))$, $i_o = S_r(j) - i_m$.

If $i = i_m$, $O_l(i) = 1$.

else $O_l(i) = 0$.

Using both geometric and photometric constraints:

For each j, when $\#(S_r(j)) > 1$

Set $i_m = \arg\max_i d(S_r(j))$, $i_o = S_r(j) - i_m$

, and $i_n = \arg\min_i E(S_r(j), d(S_r(j)))$.

If $i_m = i_n$

If $i = i_m$, $O_l(i) = 1$.

else $O_l(i) = 0$, for all $i = i_o$.

else

$O_l(i) = 0$, for all $i \in S_r(j)$.

Fig. 6. Pseudo code for asymmetric occlusion detection.

Our main goal is not to detect the occluded pixels in an image correctly but to determine a candidate set of occluded pixels. Then, reasonable cost functions in the candidate set are assigned. As many occluded pixels as possible should be included in the candidate set. Although some visible pixels may be contained in the candidate set, this problem is solved by using the proposed cost aggregation. For asymmetric occlusion detection, we use geometric and photometric constraints. To determine whether a pixel is visible or not, we have to evaluate the disparity values of the neighboring pixels. Since it is assumed that the epipolar line is parallel, we only consider the pixels in the corresponding horizontal line. In fact, to use ordering constraint means that only the right pixel of each pixel is considered. The disparity of the occluding pixels is generally larger than that of the occluded pixels. Before we define the visibility function of the pixels based on this principle, we describe the function $S_r(j)$ as a set of pixels in the right image

$$S_r(j) = \{i | i - d(i) = j, \text{ all } i \text{ with } 0 \le i \le W - 1\}$$

where i and j represent the x coordinates of the left and right images, respectively. W represents the width of the image and d represents the disparity of the pixel. Fig. 6(a) shows the pseudo code of the asymmetric occlusion detection method. O_l represents the visibility function of the left image, which takes the value 1 (or 0) when the pixel is visible (or occluded). The approaches which exploit the uniqueness constraint determine the visibility function of the reference image with the disparity fields estimated from the other image when there are multiple matching points at pixels in the other image. However, since

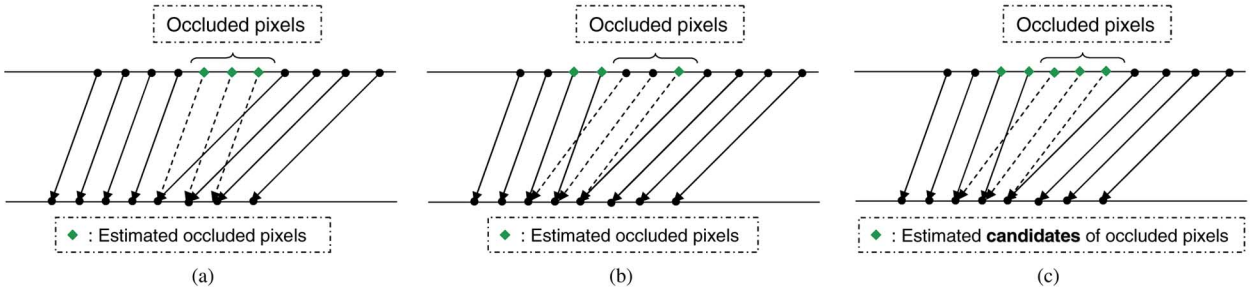


Fig. 7. Several cases of asymmetric occlusion detection: (a), (b) using the geometric constraints only, (c) using both the geometric and photometric constraints.

the proposed method uses the disparity fields on the reference image only, the pixel with the largest disparity among a set of multiple matching pixels is considered as visible and the remaining pixels are considered occluded. This is valid only if the occluding pixels have reliable disparities. Fig. 7 shows several cases of asymmetric occlusion detection. As shown in Fig. 7(a), if the disparities in the occluded pixels are smaller than those of the visible pixels, we are able to accurately detect the occluded region. Otherwise, the occluded pixels block the other visible pixels, as shown in Fig. 7(b). We use the photometric constraint to evaluate the reliability of the occluding pixels. We determine a set of occlusion candidates instead of a set of occlusions by using this constraint. This is a trivial problem since the cost function in the pixels is estimated with reliable neighboring pixels using the proposed cost aggregation scheme. Fig. 6(b) shows occlusion detection by using both geometric and photometric constraints. E is the cost function computed by (13). The costs at the occluded pixels are generally larger than those of the visible pixels. If the cost at the pixel, which is determined as occluding pixels by geometric constraints, is not smaller than that of the remaining occluded pixels, we can not guarantee the reliability of the occluding pixels. Therefore, all the pixels in $S_r(j)$ are used as occlusion candidates, as shown in Fig. 7(c). By using the visibility function O_l, we redefine the iterative scheme in (13) with (18), shown at the bottom of the page.

Fig. 8 shows the overall process of the proposed occlusion handling method. When the 3-D cost volume is given, we are able to estimate the disparity by using several techniques as an optimization method such as the WTA method, dynamic programming, belief propagation, graph cut, and so on. In this paper, we only use the WTA method to evaluate performance of proposed cost aggregation. However, other optimization techniques (for instance, belief propagation) can also be used in the proposed cost aggregation scheme. Most conventional methods perform the cost aggregation and optimization steps independently, in other words, after the cost aggregation step, the optimization step is performed with the aggregated cost function.

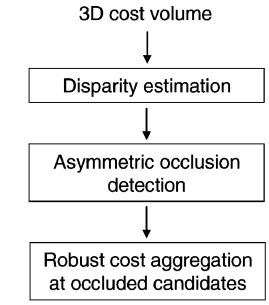


Fig. 8. Occlusion handling process: given the 3-D cost volume, we assign reasonable data to the occluded candidates. In this case, various techniques can be used as the optimization method for disparity estimation.

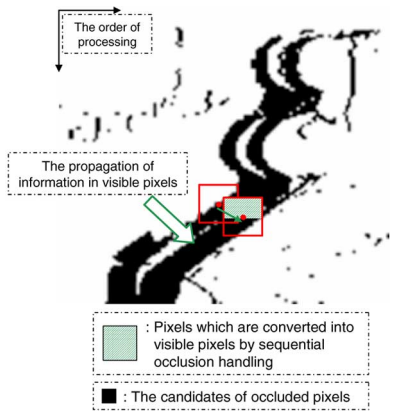


Fig. 9. Propagation of information at the visible pixels in sequential occlusion handling.

However, we perform the cost aggregation and optimization steps interactively. If we use belief propagation for disparity estimation at each level, the message computed at each specific level can be used to initialize the message of the finer level. This scheme is very similar to hierarchical belief propagation [36].

Occlusion handling is sequentially performed. The order of the iterative scheme is generally from the upper left to the lower right parts of an image, as previously mentioned. After the cost

$$E^{k+1}(p) = \frac{O_l(p)e(p) + \lambda \sum_{m \in N_c(p)} O_l(m)w(p,m)E^{k+1}(m) + \lambda \sum_{m \in N_n(p)} O_l(m)w(p,m)E^{k}(m)}{O_l(p) + \lambda \sum_{m \in N(p)} O_l(m)w(p,m)} \tag{18}$$

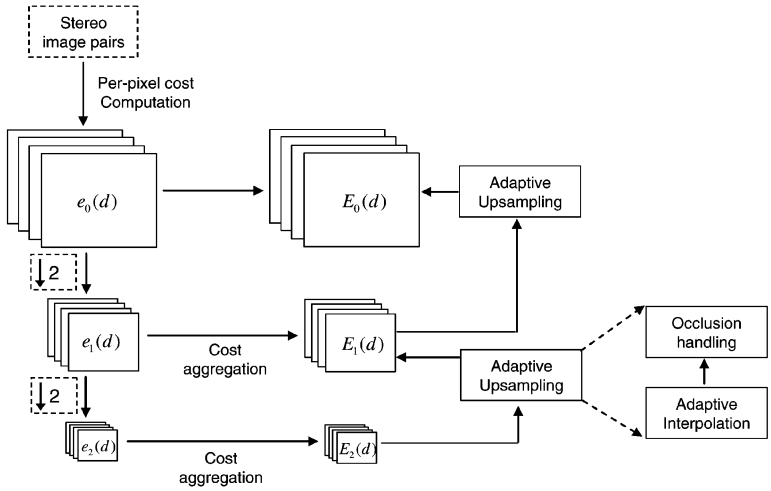


Fig. 10. Overall framework of the proposed method. Occlusion handling is done once at each level. Cost aggregation is not done at the finest level, only occlusion handling is performed.

aggregation scheme is performed with the neighboring visible pixels at the pixels of the set of occlusion candidates, the pixels become visible, in other words, $O_l(p) = 1$. This is very reasonable for occlusion handling since the occluded pixels are generally on the left side of the visible pixels and the occluded pixels aggregated with the visible pixels are used as visible pixels again in (13). Fig. 9 shows the process of sequential occlusion handling. The information of the visible pixels is propagated to estimate the cost function at the occluded pixels by using the proposed occlusion handling.

The proposed occlusion handling method is different from the extrapolation technique widely used for occlusion handling. While the extrapolation technique is just filling by using the disparities of the visible pixels, the proposed method propagates the information of the visible pixels into that of the occluded pixels. This is very similar to the concept of edge-preserving nonlinear diffusion.

V. Experimental Results

A. Overall Framework and Experimental Environments

The basic framework of the proposed method is to perform cost aggregation in a coarse-to-fine manner. Fig. 10 shows the overall process of the proposed method. The cost function on the coarsest level is computed with the iterative method in (13). In order to initialize the cost function in the finer level, adaptive interpolation is performed with (16), and then occlusion handling is performed (once at each level). This process is repeated until the finest level is reached.

We evaluate the performance of the proposed method and compared it with state-of-the-art methods in the Middlebury test bed [38]. We use the following test data sets: "Tsukuba," "Venus," "Teddy," and "Cone." The results for each test dataset are evaluated by measuring the percentage of bad matching pixels (where the absolute disparity error is greater than 1 pixel). The measurement is computed for three subsets of an image: nonocc (the pixels in the nonoccluded regions), all (the pixels in both the nonoccluded and half-occluded regions), and disc (the visible pixels near the occluded regions).

The proposed method is tested using the same parameters for all the test images. The two parameters in the weighting function are $r_c = 8.0$, $r_s = 8.0$, and the weighting factor is $\lambda = 1.0$. We use the multiscale approach at four levels, and the number of iterations is $(3, 2, 2, \times)$, on a coarse to fine scale. The iteration number of the finest level is not defined since we use the adaptive interpolation technique in the up-sampling step, as mentioned in Section III. The sizes of the sets of neighbor pixels are 5×5, 7×7, 9×9, and 9×9. In the finest level, only occlusion handling is performed.

B. Performance Analysis

Fig. 11 shows the results of the proposed method for the test bed images. We obtained the final disparity maps with a sub-pixel estimation using quadratic fitting of the estimated cost functions. The proposed method yielded accurate results for the discontinuity, occluded, and textureless regions. Table I shows objective evaluation results for comparison with other state-of-the-art methods. The results show that the proposed method obtained comparable performance with state-of-the-art methods, and the best performance among many cost aggregation techniques. We used only a simple local optimization method (WTA) with no color segmentation, while most state-of-the-art methods use various cues in low level vision such as color segmentation and global optimization techniques such as graph cut and belief propagation.

Fig. 12 shows the results obtained by the proposed occlusion handling method. Given the aggregated cost function, the occlusion candidate map was estimated with the geometric and photometric constraints. The disparity map was computed with the WTA method. The occlusion candidate set contained as many occluded pixels as possible in order to perform occlusion handling well. In Fig. 12(b), although some visible pixels were in the occlusion candidate map, the costs were accurately computed with neighboring reliable visible pixels. The proposed occlusion handling method was sequentially performed so that pixels converted into visible pixels with sequential occlusion

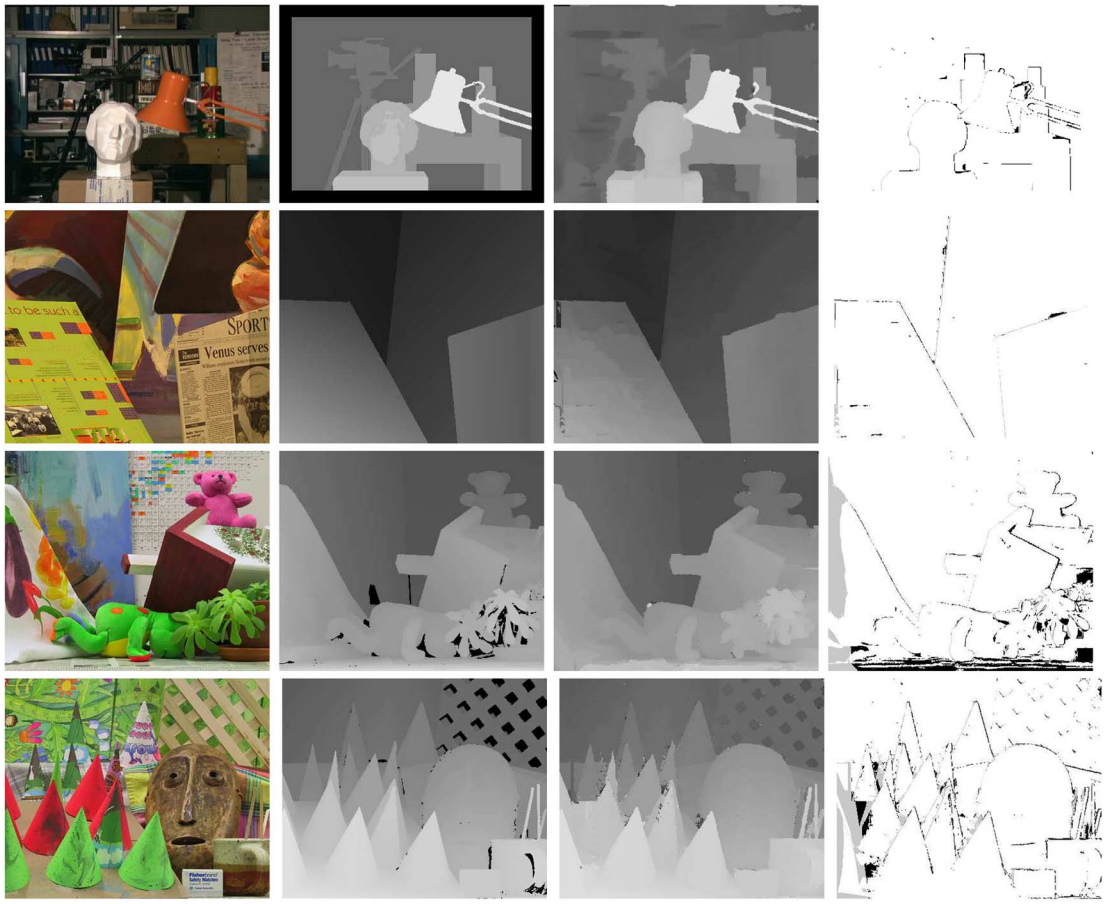


Fig. 11. Results for (from top to bottom) "Tsukuba," "Venus," "Teddy," and "Cone" image pairs: (a) original images, (b) ground truth maps, (c) our results, (d) error maps.

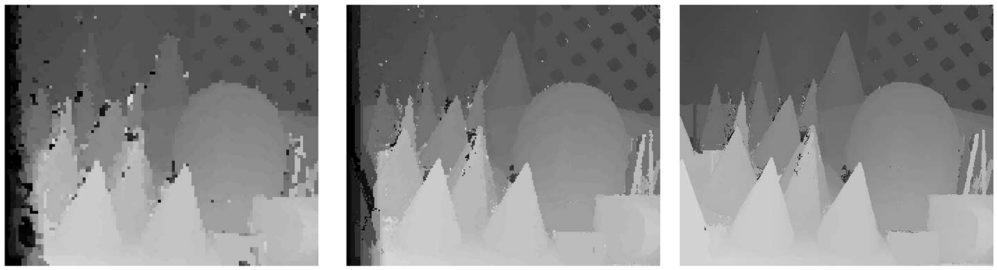
TABLE I
OBJECTIVE EVALUATION FOR THE PROPOSED METHOD WITH THE MIDDLEBURY TEST BED

Algorithm	*Tsukuba*			*Venus*			*Teddy*			*Cone*		
	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc
AdaptingBP [27]	1.11	1.37	5.79	0.10	0.21	1.44	4.22	7.06	11.8	2.48	7.92	7.32
DoubleBP [28]	0.88	1.29	4.76	0.14	0.60	2.00	3.55	8.71	9.70	2.90	9.24	7.80
SymBP+occ [18]	0.97	1.75	5.09	0.16	0.33	2.19	6.47	10.7	17.0	4.79	10.7	10.9
Segm+visib [29]	1.30	1.57	6.92	0.79	1.06	6.76	5.00	6.54	12.3	3.72	8.62	10.2
C-SemiGlob [30]	2.61	3.29	9.89	0.25	0.57	3.24	5.14	11.8	13.0	2.77	8.35	8.20
DistinctSM [31]	1.21	1.75	6.39	0.35	0.69	2.63	7.45	13.0	18.1	3.91	9.91	8.32
Our method	**1.38**	**1.96**	**7.14**	**0.44**	**1.13**	**4.87**	**6.80**	**11.9**	**17.3**	**3.60**	**8.57**	**9.36**
OverSegmBP [32]	1.69	1.97	8.47	0.50	0.68	4.69	6.74	11.9	15.8	3.19	8.81	8.89
RegionTreeDP [33]	1.39	1.64	6.85	0.22	0.57	1.93	7.42	11.9	16.8	6.31	11.9	11.8
EnhancedBP [34]	0.94	1.74	5.05	0.35	0.86	4.34	8.11	13.3	18.5	5.09	11.1	11.0
AdaptWeight [14]	1.38	1.85	6.90	0.71	1.19	6.13	7.88	13.3	18.6	3.97	9.79	8.26
SegTreeDP [35]	2.21	2.76	10.3	0.46	0.60	2.44	9.58	15.2	18.4	3.23	7.86	8.83
Layered	1.57	1.87	8.28	1.34	1.85	6.85	8.64	14.3	18.5	6.59	14.7	14.4
GC+occ	1.19	2.01	6.24	1.64	2.19	6.75	11.2	17.4	19.8	5.36	12.4	13.0
MutiCamGC	1.27	1.99	6.48	2.79	3.13	3.60	12.0	17.6	22.0	4.89	11.8	12.1
TensorVoting	3.79	4.79	8.86	1.23	1.88	11.5	9.76	17.0	24.0	4.38	11.4	12.2
GenModel	2.57	4.74	13.0	1.72	3.08	16.9	6.86	15.0	19.2	4.64	14.9	11.4

handling were also used for occlusion handling for other occluded pixels. Occluded pixels generally consist of two categories. One is generated by a front object and appears in the right side of the background. The other appears in the left border region of the image. Thus, we performed occlusion handling from the right to the left side at first. The starting point of the right side was equivalent to the search range. Then, we performed occlusion handling from the left to the right. By propagating the cost values of reliable pixels, we were able to assign reasonable data to the occluded pixels. The weighting function w played the role of discontinuity localization, such as in anisotropic diffusion. We used an asymmetric weighting function that was slightly different from that explained in (12). The asymmetric weighting function computed with the left image was used for occlusion handling since the symmetric weighting function is valid only at the visible pixels.

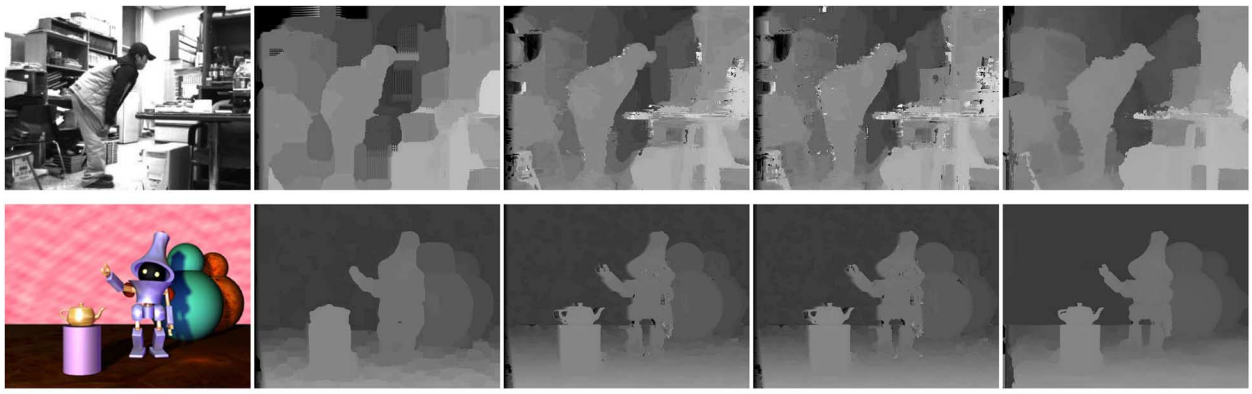
(a) (b) (c)

Fig. 12. Results in occlusion handling: (a) disparity map before occlusion handling, (b) occlusion candidate, (c) disparity map after occlusion handling. The proposed occlusion handling method is different from the extrapolation technique widely used for occlusion handling.



(a) (b) (c)

Fig. 13. Disparity maps in each level in a multiscale approach: (a) level 3, (b) level 2, (c) level 1. By using adaptive interpolation in the up-sampling procedure, we were able to acquire the disparity map on the next level with the cost aggregation on the current level.



(a) (b) (c) (d) (e)

Fig. 14. Results for (from top to bottom) captured image pairs ("Boy" and "Robot"): (a) original images, (b) shiftable window (c) adaptive weight window, (d) two-pass approximation method, (e) our method. The processing time of the proposed method is nearly 10% of that of the adaptive weight approach, as shown in Table II.

TABLE II
PROCESSING TIMES FOR CAPTURED IMAGE PAIRS

Algorithms	Shiftable [37]	AdaptWeight [14]	2passApp [15]	Our method
'Boy'	4.1s	87.5s	6.2s	9.8s
'Robot'	19.2s	428.2s	31.5s	42.5s

Fig. 13 shows the intermediate results of the multiscale approach. Since the cost function in each level was obtained after performing adaptive interpolation, the cost function was considered as that in the finer level. In that sense, if cost aggregation was performed at level l, the cost function of level l–1 could have been obtained. We found that the estimated disparity map in level 1 had the finest resolution as shown in Fig. 13(c).

The experiment was additionally performed using other images such as "Boy" (320×240 pixels) and "Robot" (640×480 pixels). The "Boy" image, captured by the Bumblebee camera of Point Grey Research, Inc., has very complex geometry. The "Robot" image was generated by 3-D studio MAX. The search ranges for them were 30 and 35, respectively. The results for the images are shown in Fig. 14. For comparison, we also show the results of the shiftable window [37], the adaptive weight window [14], and two-pass approximation [15]. In the adaptive weight approach, the parameter setting was the same as that in [14]. We confirm in Fig. 14 that the proposed method

showed excellent performance and disparities in the occluded regions were accurately estimated. The processing times of the methods are shown in Table II. The processing time of the proposed method was nearly 10% of that of the adaptive weight approach.

VI. Conclusion

In this paper, we have proposed the cost aggregation and occlusion handling method for stereo matching with the weighted least square. By solving the iterative scheme with several acceleration techniques such as the Gauss–Seidel method, the multiscale approach, and adaptive interpolation, we efficiently estimated an accurate disparity map. The information at the visible pixels was propagated into the occluded pixels by sequential occlusion handling. This process was very similar to edge-preserving nonlinear diffusion. The experimental results show that the performance of the proposed method is comparable to state-of-the-art methods and is the best among all cost aggregation methods in the Middlebury stereo datasets. Since we did not use any color segmentationt or global optimization techniques, the proposed method was efficient. In further research, we will apply the global optimization technique such as hierarchical belief propagation to the proposed multiscale cost aggregation scheme, as mentioned in the experimental results.

Acknowledgment

The authors would like to thank the anonymous reviewers of their paper for the many helpful suggestions.

References

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, Apr. 2002.

[2] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: Theory and experiment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 9, pp. 920–932, Sep. 1994.

[3] Y. Boykov, O. Veksler, and R. Zabih, "A variable window approach to early vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1283–1294, Dec. 1998.

[4] O. Veksler, "Stereo correspondence with compact windows via minimum ratio cycle," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1654–1660, Dec. 2002.

[5] O. Veksler, "Fast variable window for stereo correspondence using integral images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, vol. 1, pp. 556–561.

[6] Y. Zhang and C. Kambhamettu, "Stereo matching with segmentation-based cooperation," in *Proc. Eur. Conf. Computer Vision*, 2002, pp. 556–571.

[7] H. Tao and H. S. Sawhney, "Global matching criterion and color segmentation based stereo," in *Proc. IEEE Workshop on the Application of Computer Vision*, 2000, pp. 246–253.

[8] H. Lim and H. Park, "A dense disparity estimation method using color segmentation and energy minimization," in *Proc. IEEE Int. Conf. Image Processing*, 2006, pp. 1033–1036.

[9] M. Gong and R. Yang, "Image-gradient-guided real-time stereo on graphics hardware," in *Proc. IEEE 3DIM*, 2005, pp. 548–555.

[10] L. Tang, C. Wu, and Z. Chen, "Image dense matching based on region growth with adaptive window," *Pattern Recognit. Lett.*, vol. 23, pp. 1169–1178, 2002.

[11] S. Yoon, D. Min, and K. Sohn, "Fast dense stereo matching using adaptive window in hierarchical framework," in *Proc. Int. Symp. Visual Computing*, 2006, pp. 316–325.

[12] A. Fusiello, V. Roberto, and E. Trucco, "Efficient stereo with multiple windowing," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 858–863.

[13] M. Okutomi, Y. Katayama, and S. Oka, "A simple stereo algorithm to recover precise object boundaries and smooth surfaces," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 261–273, 2002.

[14] K. Yoon and I. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.

[15] L. Wang, M. Liao, M. Gong, R. Yang, and D. Nister, "High-quality real-time stereo using adaptive cost aggregation and dynamic programming," in *Proc. IEEE 3DPVT*, 2006, pp. 798–805.

[16] A. Bobick and S. Intille, "Large occlusion stereo," *Int. J. Comput. Vis.*, vol. 33, no. 3, pp. 1–20, Sep. 1999.

[17] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. IEEE Int. Conf. Computer Vision*, 2001, pp. 508–515.

[18] J. Sun, Y. Li, S. Kang, and H. Shum, "Symmetric stereo matching for occlusion handling," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 399–406.

[19] L. Alvarez, R. Deriche, J. Sanchez, and J. Weickert, "Dense disparity map estimation respecting image discontinuities: A PDE and scale-space based approach," *J. Vis. Commun. Image Represent.*, vol. 13, pp. 3–21, 2002.

[20] H. Kim, Y. Choe, and K. Sohn, "Disparity estimation using region-dividing technique with energy-based regularization," *Opt. Eng.*, vol. 43, no. 8, pp. 1882–1890, Aug. 2004.

[21] D. Scharstein and R. Sezliski, "Stereo matching with nonlinear diffusion," *Int. J. Comput. Vis.*, vol. 28, no. 2, pp. 155–174, June 1998.

[22] M. Heath, *Scientific Computing: An Introductory Survey*. New York: McGraw Hill, 2002.

[23] A. Willsky, "Multiresolution markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396–1458, Aug. 2002.

[24] G. Meerbergen, M. Vergauwen, M. Pollefeys, and L. Van, "Gool A hierarchical symmetric stereo algorithm using dynamic programming," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 275–285, 2002.

[25] R. Yang and M. Pollefeys, "Multiresolution real-time stereo on commodity graphics hardware," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, pp. 1211–1217.

[26] S. Geman and D. E. McClure, "Bayesian image analysis: An application to single photon emission tomography," *Proc. Statist. Comput. Sect.*, pp. 12–18, 1985.

[27] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2006, pp. 15–18.

[28] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, "Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 2347–2354.

[29] M. Bleyer and M. Gelautz, "A layered stereo algorithm using image segmentation and global visibility constraints," in *Proc. IEEE Int. Conf. Image Processing*, 2004, pp. 2997–3000.

[30] H. Hirschmueller, "Stereo vision in structured environments by consistent semi-global matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 2386–2393.

[31] K. Yoon and I. Kweon, "Stereo matching with the distinctive similarity measure," presented at the IEEE Int. Conf. Computer Vision, 2007.

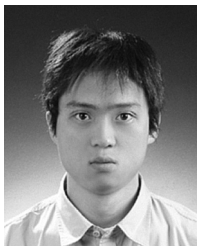
[32] L. Zitnick and S. B. Kang, "Stereo for image-based rendering using image oversegmentation," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 49–65, 2007.

[33] C. Lei, J. Selzer, and Y. Yang, "Region-tree based stereo using dynamic programming optimization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 2378–2385.

[34] S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs, "Temporally consistent reconstruction from multiple video streams using enhanced belief propagation," presented at the IEEE Int. Conf. Computer Vision, 2007.

[35] Y. Deng and X. Lin, "A fast line segment based dense stereo algorithm using tree dynamic programming," in *Proc. Eur. Conf. Computer Vision*, 2006, pp. 201–212.

[36] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, pp. 261–268.
[37] S. B. Kang, R. Szeliski, and C. Jinxjang, "Handling occlusions in dense multi-view stereo," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, pp. 103–110.
[38] [Online]. Available: http://vision.middlebury.edu/stereo



Dongbo Min (S'07) received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2003 and 2005, respectively. He is currently pursuing the Ph.D. degree at Yonsei University.

His research interests include stereo vision, 3-D modeling, view synthesis, and hybrid sensor system.

Kwanghoon Sohn (M'92) received the B.E. degree in electronics engineering from Yonsei University, Seoul, Korea in 1983, the MSEE degree in electrical engineering from University of Minnesota, Minneapolis, in 1985, and the Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, in 1992, respectively.

He was a senior member of research staff at the Electronics and Telecommunications Research Institute (ETRI), Korea, from 1992 to 1993. Also, he was a Postdoctoral Fellow with the Medical School of Georgetown University, Washington, DC. He was a Visiting Professor at Nanyang Technological University from 2002 to 2003. He is currently a Professor in the School of Electrical and Electronic Engineering, Yonsei University. His research interests include 3-D image processing, computer vision, and image communication.

Dr. Sohn is a member of SPIE.