IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received April 27, 2020, accepted May 18, 2020, date of publication May 25, 2020, date of current version June 8, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2997073*

# iGG-MBS: Iterative Guided-Gaussian Multi-Baseline Stereo Matching

**PATHUM RATHNAYAKA**[1] **AND SOON-YONG PARK**[2]

[1]School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, South Korea
[2]School of Electronics Engineering, College of IT Engineering, Kyungpook National University, Daegu 41566, South Korea

Corresponding author: Soon-Yong Park (sypark@knu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT: Ministry of Science and ICT) under Grant 2018M2A8A5083266, the BK21 Plus project funded by the Ministry of Education, South Korea, under Grant 21A20131600011, and in part by the Cross-Ministry Giga KOREA Project grant funded by the Korea government (MSIT), Real-time 4D reconstruction of dynamic objects for ultra-realistic service, under Grant GK17P0300.

**ABSTRACT** This paper presents an improved dense disparity estimating technique for a collection of multi-baseline stereo (referred to as *MBS* in the text) images. The flow of the proposed system consists of two main frameworks: a preliminary cost calculation and initial disparity estimating framework, and an iterative cost refinement framework. The first framework implements an accurate multi-baseline stereo cost (referred to as *MBSC* in the text) calculation method, and a scan line optimization inspired by the Semi Global Matching (*SGM*) algorithm. Cost volumes of each two-view camera pair are calculated by fusing two pixel dissimilarity measures: *i)* weighted Census transformation and *ii)* sum of absolute difference color consistency term (*SAD-Census*). The initial disparity map between reference and the matching view with the largest baseline displacement is calculated by summing-and-interpolating SAD-Census costs of the current and all neighboring camera pairs in-between, and taking the minimum after aggregating for sixteen directions. The second framework refines the aggregated MBSC volume recursively. In each iteration, individual pair-wise disparity maps are used to warp matching views towards the reference to create binary masks that resemble overlapping differences. White locations in the mask represent incorrect correspondence matches, thus a penalty is added for costs associated with, adapting a Gaussian modulating function. This significantly reduces the selection probability of incorrect disparity minima in proceeding iterations. A Guided filter-based Rolling Guidance filter is applied to further up-vote the probability of pixels with the lowest costs, which are similar or close enough to ground truth readings. Through experimental results evaluated on the Middlebury dataset, we show that our method leads to effective and efficient multi-baseline disparity estimations.

**INDEX TERMS** Stereo vision, multi-baseline, stereo matching, iterative, refinement, disparity mapping.

## I. INTRODUCTION

Computer and machine vision is defined as an interdisciplinary scientific field that is concerned with theories and techniques for developing strategies to help computers to gain a high-level understanding of real-world situations through contents of digital images or video sequences. It includes methods such as image acquisition, processing, analysis, and understanding [1], [2]. Calculating three-dimensional information from, and establishing accurate dense correspondence matching between; two or multiple image sequences play a significant role in this field [3]. Making the computers to gain an understanding of the world scenery observed through naked human eyes and graphically restructuring them based on contents, shapes, illumination variations, color distributions, however, is still a challenge [4]. Throughout the past few decades, many algorithms exploiting the implementations of accurate 3D information calculation and dense correspondence matching have been studied. Some of the applications benefited include simple to advanced mobile robot navigating systems, sophisticated driverless autonomous systems, augmented and virtual reality applications, teleconferencing, view synthesis, and many more [5].

In general, these proposed algorithms can be categorized into two distinctive groups as *active* and *passive*. Active methods project a light source (e.g., LED, lasers) onto the

The associate editor coordinating the review of this manuscript and approving it for publication was You Yang.

scanning area. 3D information is obtained by either scanning light spots or line stripes expanded from a spot using a cylindrical lens, or by projecting light patterns of dots or lines [6]. Notwithstanding their comprising results, they show less practicability in situations when modeling distant or fast-moving objects [7], [8] or behave poorly in outdoor environments (e.g., structured light) [9].

Passive methods, on the other hand, do not project any radiation onto the scene, but instead, measure the visible radiation that already exists on the scene surface (the ambient light reflected by the target [7]) using cameras. Modern passive systems have more diversified designs: ranging from the simplest monocular camera; to two-view; to more complicated multi-baseline cameras. However, calculating 3D information using a single camera is infeasible as they fail to determine the true scale factor [9]. Therefore, many systems – as similar to human stereopsis – use two cameras with already known relative position information, which in such the dense correspondence task becomes the well-known stereo matching problem [10].

Stereo matching denotes the problem of finding pixel correspondences in image sequences that correlate with the same 3D point in the world view. This is also known as disparity estimation, which produces a parallax image specifying relative displacements between pixels [11]. The geometry that correlates 3D world objects with their 2D projections is known as the epipolar geometry [12]. According to the taxonomy of Scharstein and Szeliski in [13], most stereo matching methods consist of four distinguishable steps: matching cost computation, cost aggregation, disparity computation/optimization, and disparity refinement.

The matching cost measures the similarity and dissimilarity between two locations defined either locally or over a supportive region [11]. This computation generally exploits the absolute, squared, or sampling insensitive differences of intensities of colors [14] to generate a 3D cost volume known as disparity space image [15]. However, most of these traditional costs are sensitive to radiometric differences, such as excessive exposure or global illumination differences. Therefore, costs based on non-parametric transformations such as Census or costs combined with image gradients such as mutual information are adapted. Since the initial matching cost volume is sensitive and noisy, it is aggregated in a supportive region [16]. This enforces the piece-wise coherency of a resultant disparity map [10]. In disparity computation, disparity related to the lowest matching costs are selected according to the winner takes all concept (WTA) [17]–[19]. Disparity refinement ensures visibility enhances by removing peaks [20] and occluded regions, consistency improvements by interpolating gaps [17], [21], and accuracy improvements based on sub-pixel interpolation [21], [22]. Once the disparity image is obtained, the depth of individual pixels can be computed using triangulation [23].

A vast collection of stereo matching algorithms for disparity mapping have been proposed throughout the past few years. Articles cited in [2] and [24] summarize few of such algorithms. Notwithstanding most of the state-of-the-art methods concentrate on two-view stereo, being limited to a single fixed baseline could cause occlusion problems. In addition, they suffer from wrong depth estimates caused by local minima in the matching cost functions [25]. It is generally recognized that using more than two views has the potential to improve the quality of depth estimation [26].

Therefore, multiview stereo, also known as multi-baseline stereo [27] (MBS according to our notation), is originated as a natural improvement to the state-of-the-art two-view case [28]. For the past few years, many MBS matching algorithms with different system configurations have been proposed. Among them, some have proposed improvements into matching cost volumes, whereas some have proposed improvements into cost aggregation. There are methods that use multiple arrays of cameras for multiview image acquisition while exploiting the advantage of silhouettes for dense mapping and proper surface reconstructions [29]. Some methods use mono cameras to capture multiview images by moving in-and-around a scanning area.

The motivation of this research paper is to propose an easy, yet robust MBS matching algorithm for dense disparity estimation. In this article, we deliberately limit our discussion to disparity mapping, and do not dive deep into 3D modeling or meshing. We refer the reader to research works in [30]–[33] for more detailed discussions about some of the currently available robust 3D modeling and meshing techniques. Having said that, we try to emphasize how our proposed method can be extended and be easily adopted into those researches for accurate dense disparity mappings even with a less number of images.

The structure of our paper is as follows: In Section *II*, we briefly discuss few of the MBS matching techniques currently available and try to point-out their advantages as well as disadvantages. In Section *III*, we describe the preliminaries that we require for our proposed disparity mapping technique. First, we discuss how two-view SGM algorithm [34], [35] can be extended into MBS platform. Then we talk about the calculation of cost volumes by looking into the perspectives of two-view matching. This includes an in-detailed summary of our SAD and weighted Census cost calculation, and their fusion. Section *IV* presents the core of this paper: our proposed MBS disparity mapping algorithm. For simplicity, we have divided the section into two parts. The first part shows how two-view cost volumes that we described in the previous section can be extended into multi-baseline platform for an initial disparity estimation. A qualitative result analysis between the initial disparity images of ours and original MBS-SGM is also included. The second part gives an in-detailed description to our iterative cost refinement framework. In this, we talk about how the initial cost volumes can be refined considering a Gaussian modulating function and applying a rolling-guidance-based guided filter. We discuss how this approach improves the quality of disparity results w.r.t previous iteration by asserting a qualitative result analysis. A few of the post processing techniques we used for disparity refinement, such as hole-filling and filtering are also described in simple point form. Section *V* summarizes our experimental results. We performed our experiments using 12 image sets available in the Middlebury dataset. There, we make some qualitative analyses by comparing non occlusion results between ours and the original MBS-SGM algorithm. Some quantitative analyses done by comparing disparity errors in the Middlebury evaluation site are also summarized. Finally, we conclude the paper by stating our thoughts and future improvements in Section *VI*.

## II. PREVIOUS WORKS

A series of MBS matching techniques have been proposed throughout the past few years. An early research on MBS was introduced in article [36]. In this work, a linearly arranged camera setup with parallel optical axes was designed to experiment for the advantages of using both narrow and wide baselines. They pointed out that a shorter baseline results in less precision whereas a wider baseline results in higher error rates due to ambiguous false matches. A window-based matching was performed in supportive regions by computing sum of squared difference values between each two-view stereo pair, and were summed together into an inverse distance representation, which they called as the *SSSD-in-inverse-distance (Sum of Sum of Squared Difference-in-inverse-distance)*. As of drawbacks, the algorithm showed some limitations in occlusion modeling.

The same SSSD cost volume was used in the work of Kang *et al.* [37]. In this work, they paid more concerns on mitigating occlusion handling problems. A combination of shiftable windows and a dynamically selected subset of neighbor images were used to calculate the SSSD matching cost volume. If a pixel in the reference view was supposed to be occluded, a minimum matching cost between a subset of the cameras was selected and used as the respective cost value. However, since it is difficult for vision-based methods to reconstruct a 3D model with high accuracy, the approach is not suitable for 3D modeling of wide area outdoor environments [38].

A variable multi-baseline, multi-resolution stereo matching technique was presented by Gallup *et al.* [39]. They discussed the importance of exploiting both wide and narrow baselines for far and near range acquisitions while adopting the fundamentals of plane-sweeping stereo. Though they managed to illustrate the benefits of fine grain data association strategies in multiview stereo, some argue that the method cannot be easily generalized into irregularly captured datasets [32].

Li *et al.*, proposed a long baseline global stereo approach based upon short baseline estimations [40]. Their main discussion relied on introducing a novel idea that tends to improve both efficiency and accuracy in wide baseline global stereo matching. A relationship between disparities of a corresponding point in between different baselines was proposed by considering quantized error where the disparity search range under the long baseline was reduced by guidance of the short baseline for efficiency improvements. However, the method shows lack of improvements in depth discontinuity handling.

A multiview stereo approach for airborne disparity mapping was described in [41]. They tried to emphasize that dense image matching algorithms based on SGM are capable of successfully matching more than 99% of all pixels and of providing higher matching accuracy rates. In order to solve the increment in redundancy caused due to using multiple images, they defined a central image as the base image and matched it against the surrounding images according to an overlapping probability. Density drops of disparity mappings due to outlier filtering are solved by combining several stereo image pairs for multiple parallax estimation. As for accuracy analyses, they implemented both point-to-point and point-to-plane distance estimations, which however, are much sensitive to blunders and random noise within the dense matching point clouds. In addition, the quality measures are less reliable or persuasive if calculated without consideration of the breaklines in natural scenes (e.g., edges and bumpy terrain) [42].

An embedded MBS approach for real-time application using a four narrow MBS camera setup with on-board FPGA was described in [25]. The advantage of this design is its light weight, which makes it more appropriate to fit in most mobile robot systems. However, their census-based local stereo matching approach has shown limited accuracy levels compared to most global and/or semi-global approaches.

Poggi *et al.*, introduced a framework aimed at enforcing a trinocular assumption for training a CNN network in an unsupervised manner for monocular depth estimations [26]. They named this network as the *3Net*. The trinocular assumption they defined mitigates most of the limitations caused by using binocular stereo images as supervision. The interleaved training protocol outperformed most unsupervised techniques exist, ensuring its position as a state-of-the-art method. However, being limited to offline training makes it bit inconvenient to be used as an online adaptation to unseen environments.

## III. RESEARCH PRELIMINARIES

In this section, we try to draw the attention of the reader toward the two-view SGM algorithm, how this method can be extended into multiview matching, along with some mathematical description to our proposed cost volume calculation. This we did to mitigate any possible confusions that could rise when discussing MBSC fusion in proceeding sections.

### A. SGM FOR MBS MATCHING

According to the taxonomy of stereo matching, the base of any matching algorithm contains a matching cost volume – a measurement of the similarity and dissimilarity of corresponding locations between two or multiple images – which is aggregated to remove matching ambiguities. The stereo version of SGM uses a pixel-wise Mutual Information (MI)

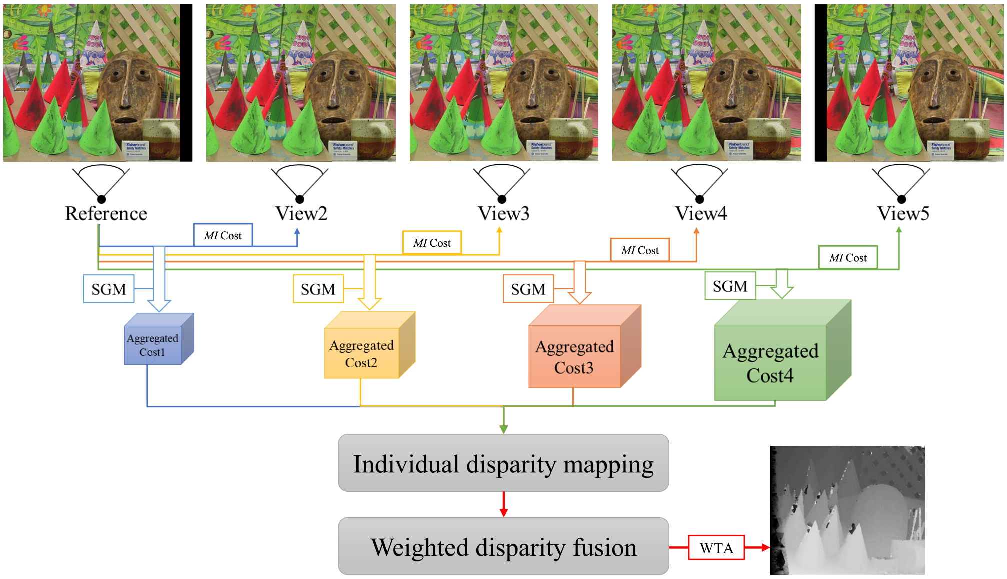


**FIGURE 1.** Conventional SGM-based MBS disparity mapping. Individual disparity pairs are weighted summed according to their baseline ratios to generate the raw/unfiltered disparity result.

that is calculated in a hierarchical manner to generate their matching cost volume.

A global energy term is approximated which is solved as a one-dimensional minimization problem along several 1D independent paths [43], where in each path, pixel costs are aggregated considering neighboring pixels and additional smoothing penalties.

The process is easily extended into MBS matching, of which pair-wise matching between the reference and each adjoining view are performed individually. Instead of calculating a combined pixel-wise cost volume between reference and matching views, separate disparity results of each two-view pair are weighted summed considering individual scalings. However, scaling parameters should be linear to the length of baselines between cameras and all images must project onto a common plane that has the same distance to all optical centers (see *multibaseline matching* in [35]). Disparities with lowest aggregated costs are then used to create the disparity image. Fig.1 summarizes the process of computing a left disparity result for a MBS configuration with five views. View$_1$ is considered as the reference, followed by Views$_{2, 3, 4}$ and View$_5$, which is considered as the matching view with the largest baseline displacement with the reference.

### B. TWO-VIEW SAD, CENSUS COST VOLUME CALCULATION AND FUSION

As mentioned, the original SGM algorithm considers a hierarchical MI cost volume for its disparity estimation. Notwithstanding their comprising results, using MI to calculate matching costs between multiple views in this similar hierarchical fashion could cause computational complexities and lead into run-time limitations. This led us to the idea of fusing two cost terms that required comparatively less computation power, but capable of providing high reliable correspondence matches. Instead of using MI, in this research; we define a SAD-Census cost volume by fusing two pixel dissimilarity measures: a sum of absolute difference-based color consistency cost ($C_{SAD}$) and a Hamming distance-based weighted color Census cost ($C_{Census}$).

Multiple exploitation of color information based on regional matches have proved to provide better performance trade-offs between computation complexities and to increase matching reliabilities by reducing ambiguities in areas where depth discontinuities exist. In a general two-view framework, the color consistency term can be defined as a regional sum of averaged absolute intensity difference of RGB channels between two compared pixel locations in the reference view and its adjoining matching view. For any given pixel $P_{(x,y)}$ in the reference view with a maximum disparity level of $d$, and $I$ representing the pixel intensity which is propagated through a regional squared window $W$, its color consistency cost can be summarized as it is shown in (1).

$$C_{SAD}(p, d) = \frac{\sum_W \left( \sum_{i=R,G,B} \left| I_i^{ref}(P_{x,y}) - I_i^{match}(P_{x-d,y}) \right| \right)}{3} \tag{1}$$

Census transformation cost, on the other hand, encodes local structures with relative orderings of pixel intensities [44]. It limits the noise and effects of radiometric differences such as excessive exposure and global illumination variances. Also it provides high reliable matching capabilities as MI, but only requires substantially less computation power [43]. In a general two-view framework, the transformation cost can be defined as the Hamming distance between two concatenated bit strings that represent intensity changes of pixels in the reference view and its matching view respectively. This is

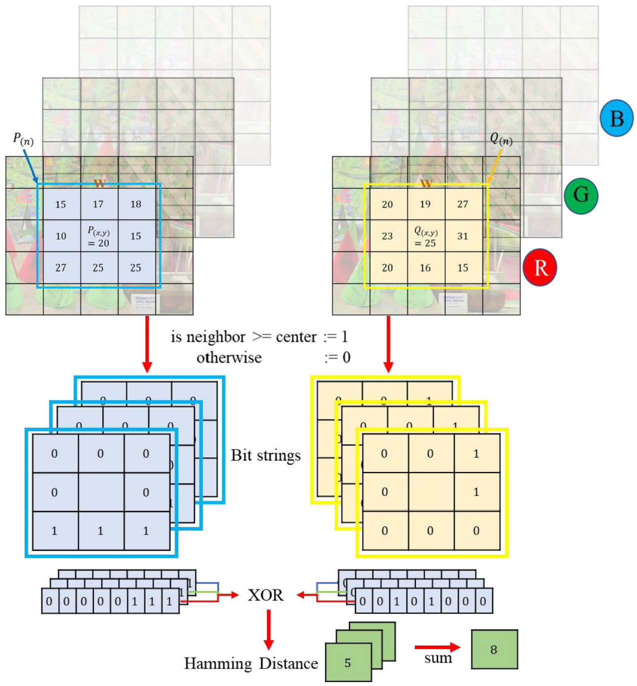


**FIGURE 2.** An example of how Census transformation between two images is calculated using a simple 3 × 3 window. Bit strings of each channel are calculated by comparing intensity changes between the center pixel and its neighborhood. Hamming distances between each respective bit string are estimated according to an XOR operation and a pop count of 1s.

exemplified in Fig.2. $P_{(x,y)}$ and $P_{n(x,y)}$ represent the reference pixel and its neighborhood in $W$, with $Q_{(x,y)}$ and $Q_{n(x,y)}$ being the matching view pixel and its neighborhood, respectively.

The first step of computing the Census cost is to generate individual Census transformation terms between views. Each pixel's intensity values are compared against their neighborhood values and are assigned binary bits of 0s and 1s accordingly. The assigned bits are concatenated and encoded into bit strings; which are next compared using an XOR operation to calculate respective Hamming distances. Equations (2) and (3) denote the general way of calculating this term for the reference view.

$$Census_P = \bigotimes_{P_{n(x,y)} \in W} \xi(P_{x,y}, P_{n(x,y)}), \tag{2}$$

where

$$\xi(P_{x,y}, P_{n(x,y)}) = \begin{cases} 0, & P_{x,y} \geq P_{n(x,y)} \\ 1, & otherwise \end{cases} \tag{3}$$

However, in this paper, we changed this original transformation term into a weighted bit string by multiplying individual elements with an additional weight vector. We asserted the values of this vector by computing Euclidean distances between pixels of the neighborhood w.r.t their center pixel. As we realized that pixels which lie much closer to the center pixel have high correspondence reliability than of the pixels lie further, we assigned higher weight values for the nearby pixels and comparatively lower weight values for the pixels in the boundary. This modified Census transformation for the blue channel is shown in Fig.3, whereas the calculation of the weight vector is defined in (4) with $\alpha$ representing an arbitrary constant parameter used to normalize the range into [0,1]. The Euclidean distance is defined as in (5).

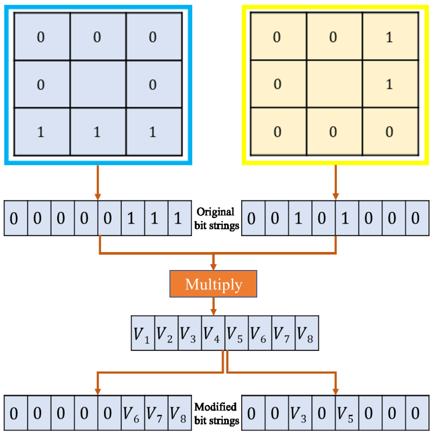


**FIGURE 3.** Multiplying bit strings with the weights vector. The weights are selected according to the Euclidean distance of pixels in the neighborhood with the center pixel in the window. Nearby pixels are given higher weights and far pixels are given lower weights.

$$V_{(P;P_n)} = 1 - \alpha \cdot \triangle_{euc}(P; P_n) \tag{4}$$

$$\triangle_{Euc}(P; P_n) = \sqrt{(P_{n(y)} - P_y)^2 + (P_{n(x)} - P_x)^2} \tag{5}$$

By combining (2) with (4) for both reference and matching views separately, we can compute the Census cost volume: $C_{\text{Census}}$ as shown in (6).

$$C_{Census}(P, d) = \sum_{i=R,G,B} \begin{matrix} \|Ham(Census_i^{ref}(P_{x,y}) - \\ Census_i^{match}(P_{x-d,y}))\| \end{matrix} \tag{6}$$

Finally, the SAD-Census cost $C(P,d)$ is computed as (7):

$$C(P, d) = 1 - \exp\left(-\frac{C_{SAD}(P, d)}{\lambda_{SAD}}\right) + 1 - \exp\left(-\frac{C_{\text{Census}}(P, d)}{\lambda_{\text{Census}}}\right), \tag{7}$$

where $\lambda_{\text{SAD}}$ and $\lambda_{\text{Census}}$ are arbitrary threshold parameters used to remove noise and matching errors. Furthermore, the exponential function maps the two costs to the range of [0,1] before combining, so that $C_{p,d}$ will not be dominated by only one cost value.

## IV. ITERATIVE MBS MATCHING

Our proposed MBS matching technique is explained in this section. We have divided this into two sub sections. In the first sub section, we describe how we can compute a MBSC volume utilizing the cost volumes that we calculated based on the perspectives of two-view stereo, and how this can be used to generate an initial disparity result. Further, we make a qualitative result comparison between ours and the original SGM algorithm. The second sub section describes how we

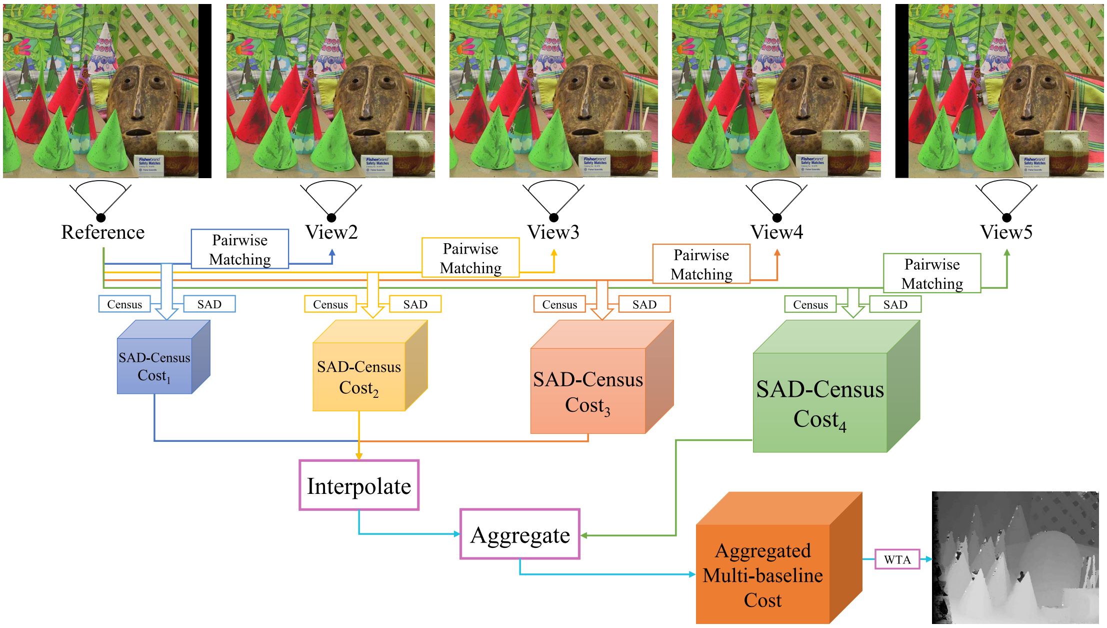


**FIGURE 4.** **Our approach to generate the initial MBS disparity image. Left most image is kept as the reference and pair-wise SAD and Census costs are calculated between this reference and all other views. Costs$_{1,2,3}$ are interpolated into the size of Cost$_4$, combined and summed for 16 directions to generate multi model aggregated cost volume. The unfiltered initial disparity map is generated by taking the minimum of the aggregated cost.**

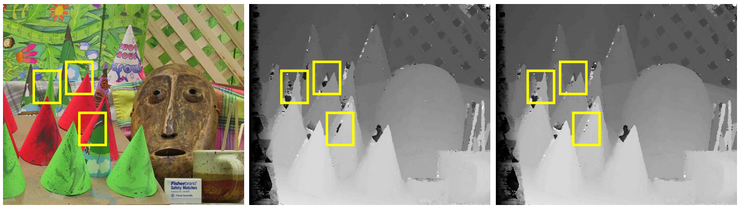


**FIGURE 5.** **Qualitative result analysis between ours and original SGM for Cones dataset. Yellow boxes depict the improvements of our approach in discontinued regions. left: reference, middle: original SGM result, right: our result.**

can refine the initial SAD-Census cost volume between each two-view for more accurate disparity estimation considering a Gaussian modulating function and a rolling guidance filter recursively. Similarly, we make a qualitative result comparison between the output of each iteration, and also with the original SGM result.

### A. MBSC VOLUME FUSION AND INITIAL DISPARITY ESTIMATION

The first framework of our proposed MBS matching technique consists of computing an initial disparity image between the reference view and the matching view that has the largest baseline displacement. We are going to exploit the two-view cost calculation method that we discussed earlier to compute a complete MBSC volume in between these views. To simplify this discussion, let us refer Fig.4. This depicts the overall process of our initial phase.

To begin with, we first calculate pair-wise SAD and Census costs between the reference and each matching view using (1) and (6), and then combine them into fused cost volumes using (7), individually. To exemplify this, if $k$ being the number of view we have including the reference ($k = 5$ in accordance with Fig.4), and $i$ representing any arbitrary *reference-matchview* pair, let us define that $C_{SAD_i}$ and $C_{Census_i}$ represent the pair-wise SAD and Census cost volumes calculated with the reference where $\{i \in k : 1 \leq i \leq k-1\}$, and $C_i$ as their fusions, respectively.

One assumption that we make; and is worth noting that; though these costs are of different sizes, but they are linearly proportional to their corresponding baselines. This literally makes an intuition that as the displacement between the reference and each matching view increases sequentially, the sizes of the cost volumes also increase accordingly. Therefore, it is necessary for scaling all the costs with different sizes to a common size, hence we interpolate all the $C_i^{lowest}$ volumes to the size of $C_i^{largest}$ volume that has the largest baseline displacement. As per our multiview system, $C_1$, $C_2$, $C_3$ denote the $C_i^{lowest}$ volumes and their interpolation is shown in (8).

$$C^{int}_{i(\{i\in k:1\leq i\leq k-1\})} = (a * C_i[\lfloor\alpha\rfloor + 1]) + (-C_i[\lfloor\alpha\rfloor] * a + C_i[\lfloor\alpha\rfloor]), \quad (8)$$

where $C[]$ represents the value of the volume to be interpolated at element index $\alpha$ which is incremented by $sizeof(C_i^{lowest}/C_i^{largest})$ range, and $a$ representing ($\alpha$-$\lfloor\alpha\rfloor$), respectively.

We combine all the $C_i^{int}$ volumes with $C_i^{largest}$ volume and recursively optimize for 16 1D-directions according to (9);

$$\begin{aligned} L_r(P,d) = C(P,d)^{multi} + min(L_r(P-r,d),\\ L_r(P-r,d\pm 1)+Pen_1,\\ \underset{i}{min}L_r(P-r,i)+Pen_2)\\ -\underset{i}{min}L_r(P-r,k) \end{aligned} \tag{9}$$

where $r$ denotes the direction, $Pen_{1,2}$ denote additional smoothness constraints, and $C(P,d)^{multi}$ denotes the full MBSC volume calculated as:

$$C(P,d)^{multi} = \sum\nolimits_{i=1,2,3} C_i^{int} \quad + C_4 \tag{10}$$

Finally, the path-wise costs are summed together in (11) to compute an aggregated MBSC volume $S(P,d)^{aggre}$, from which the final disparity for each pixel is selected based on the WTA to generate the initial MBS left disparity map $D(P)_L^{init}$ as defined in (12).

$$S(P,d)_L^{aggre} = \sum\nolimits_{r=16} L_r(P,d) \tag{11}$$

$$D(P)_L^{init} = \underset{d}{\min} S(P,d)^{aggre} \tag{12}$$

A qualitative result comparison between ours and the original MBS-SGM on the Cones dataset is summarized in Fig.5. Regions marked in yellow boxes depict the improvements of our approach in areas where depth discontinuities appear. Similarly, we can follow the same steps to generate a right disparity image $D(P)_R^{init}$ by considering view$_5$ as the reference. This we did to perform left-right-consistency check (LR-Check) in the proceeding refinement approach.

### B. ITERATIVE MATCHING COST REFINEMENT

In this sub section, we compute a new multiview aggregated cost volume by refining pair-wise cost volumes recursively. The initial disparity image that we created in our previous sub section shows comparatively good results. However, our main concern is that it still consists of outliers including holes, and mismatched regions. The main reason for these outliers to exist is that pairwise SAD and Census calculations contain erroneous matches and they are not refined properly at the aggregation step.

A simple solution for treating outliers is refining the aggregated cost volume using an edge preserving filter, such as a guided filter [45] or a bilateral filter [51] at each disparity level. This approach has been researched in many algorithms, particularly for two-view configurations [48]–[50]. Considering these concepts, we also propose an easy cost refinement approach, however, not solely being limited to edge preserving filters. Our main intention was to find a way that could reduce the selection probability of disparities related to the costs of erroneous matches at the WTA step. For this, we must increase their costs by a larger penalty value, such that only, they will be disregarded when choosing for the cost minima. However, the problem still remaining is that how we can identify which pixel we should give this penalty for? One approach is to compare the initial disparity result with sparse depth information collected from an external source, such as a LiDAR sensor, and increasing the costs of pixel indexes that deviate by a larger margin. However, integrating LiDAR sensors could be bit expensive.

Therefore, we tackle this problem by looking at it from a different perspective. Instead of integrating any additional sensors to get sparse depth data, we suggest to create 2D binary mask images that can replace them and be used to identify erroneous matches. The easiest way of creating a binary mask is calculating the absolute difference between two images. But first, either one of the images must be warped/shifted toward the view perspective of the other.

In most stereo matching techniques, the most easiest way of warping an image towards another is using an initial disparity image and shifting the pixels considering a simple lookup strategy. If the disparity result is accurate, then the warped image can be considered as a look-alike as the image that it is warped to. In general, the most common practice is warping the matching view towards the reference view. If the binary image contains more black pixels than white pixels, this emphasizes that the matching view overlaps with the reference with a few mismatches, and consequently, the disparity result is much closer to its ground truth.

Once these mismatches are properly identified, the cost values of the correlating pixel indexes can be increased by adding the penalty considering a Gaussian modulating function. However, in practice, it is difficult to obtain good results using a bad initial disparity estimation at a single iteration. Thus the process must be performed recursively, in such that the cost volume gets refined in each iteration. The disparity result of each iteration can be used as the input initial disparity to warp the matching view.

After the first phase of our proposed architecture, we have access to a reasonably good initial disparity result, which corresponds to what we believe as a reasonably good aggregated MBSC volume. Therefore, performing this cost refinement is straightforward. This iterative cost up-voting and down-voting approach is shown in Fig.6.

First, we warp the matching view$_5$ towards the view perspective of the reference using the initial LR-Check disparity image $D(P)_L^{init}$ that we created. If we consider the function $epi_{ref:match}(P,d)$ defines the epipolar line in the matching view for the reference pixel $P$ with $d$ being the line parameter, we can simply implement this warping as a general lookup in view$_5$ with $epi_{ref:match}(P,D(P)^{init})$ for all the pixels $P$. Taking $view_5^{warp}$ to resemble the warped matching image, it can be mathematically summarized as in (13).

$$view_5^{warp} = lookup[epi_{ref:match}(P,D(P)_L^{init})] \tag{13}$$

Next, we compute the absolute difference between the reference and $view_5^{warp}$ and create the binary mask image to

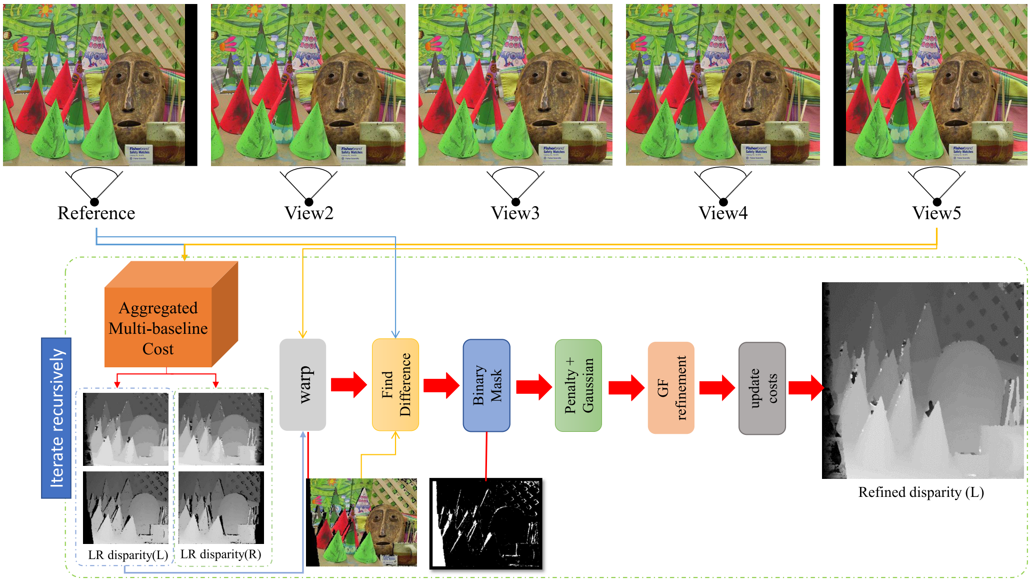


**FIGURE 6.** The proposed cost refinement approach considering reference and matching view$_5$. The initial disparity result is used to warp the matching view towards the reference, a binary mask is created to identify all the erroneously matched pixels. A higher penalty is given for the cost locations where errors appear, and the aggregated MBSC volume is further refined by adopting a rolling guidance-based guided filter.

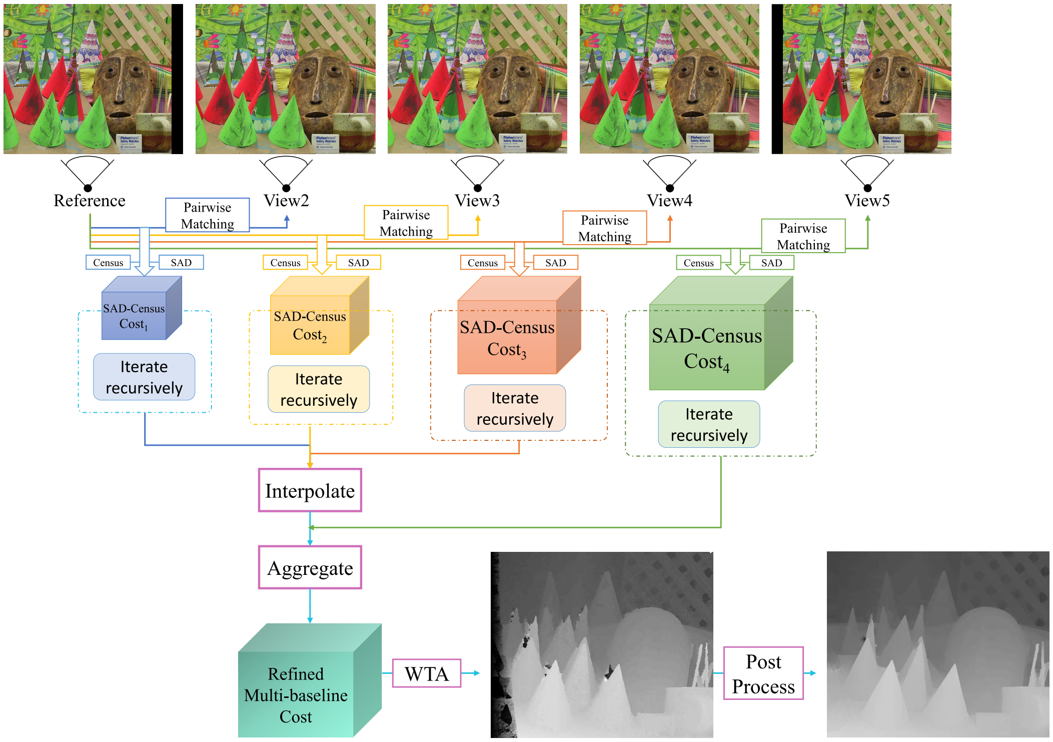


**FIGURE 7.** The proposed cost refinement approach considering all the pair-wise matching views. Each matching view is paired with the reference and processed by stereo SGM to create individual initial disparity results, which are used to create binary masks. The term ***'iterate recursively'*** denotes the method described in Fig.6. Some post processing are applied on image space to fine-tune final disparity result.

identify all the alleged correspondences. The pixels represented in white denote the erroneous matches, thus, we add the higher penalty with these costs as a convolution of a Gaussian modulating function $\otimes G(P,d)$; which is centered on the erroneous pixel. This way the convolution score corresponding to the cost $S(P,d)^{aggre}$ is multiplied by the peak of the function, while other elements are progressively lowered. If we take $S(P,d)^{aggre}$ to represent the refined aggregated

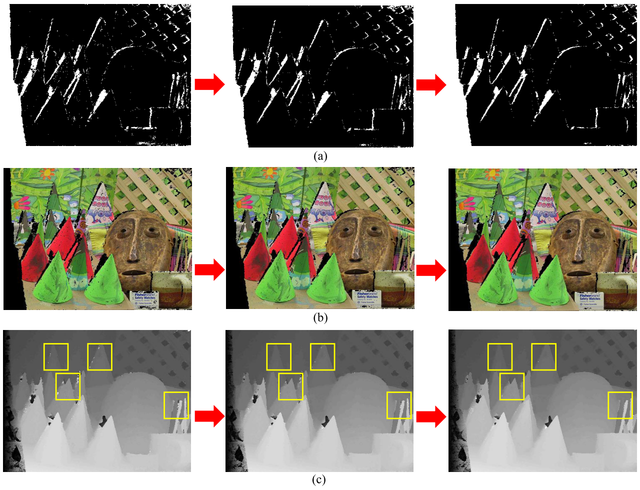


**FIGURE 8.** Results after proposed iterative aggregated MBSC refinement. (a): binary mask images for the first, second, and third iteration, (b): warped view$_5$ image towards the reference, (c): fine-tuned left disparity results.

**TABLE 1.** Values of the parameters used.

| Parameter name | Parameter value |
|---|---|
| $\lambda_{SAD}$ | 45 |
| $\lambda_{Census}$ | 65 |
| $\alpha$ | 0.3 |
| $Pen_3$ | 100 |
| $SAD_W$ | $3 \times 3$ |
| $Census_W$ | $9 \times 7$ |

cost volume and $Pen_3$ as the given penalty value, it can be mathematically represented as in (14).

$$S(P,d)^{aggre} = 1 - ([S(P,d)^{aggre} + Pen_3] \otimes G(P,d)) \quad (14)$$

In addition, we apply a guided filter [45] running on a rolling guidance [46] filtering approach to $S(P,d)^{aggre}$ volume to refine costs even further. We apply this filter at each disparity level, while choosing the reference view as the guide as:

$$S(P,d)^{aggre} = \sum_{i=min}^{max} \jmath(S(P,d_i)^{aggre} : I_{ref}) \quad (15)$$

with $\jmath$ representing the guided filter on rolling guidance, and $I_{ref}$ representing that the reference view is being used as the guide. The final processed aggregated MBSC volume is used to compute a fine-tuned raw disparity image.

We repeat this whole process for three multiple iterations. In each iteration, the disparity result computed at the previous step is used to warp the matching view$_5$ towards the reference. In our experiments, we noticed that repeating this optimization for three iterations is sufficient enough for accurate disparity estimation.

As we have evaluated our experiment results, we have witnessed that adding the penalty on erroneous pixel locations increases their matching cost values significantly, whereas the edge preserving filter further refines it, consequently reducing the overall selection probability of incorrect disparities related to the lowest matching costs in the WTA concept. However, the result still contains a few mismatches. The reason which causes these to exist is that we considered the aggregated MBSC volume directly, without separately refining individual pair-wise cost fusions used to create it.

**TABLE 2.** Quantitative evaluation of the proposed method compared with few of the algorithms on the Middlebury website for threshold 1.0.

| Algorithm | Cones | Teddy | Venus | Avg bad-pixel |
|---|---|---|---|---|
| IGSM [52] | 2.14 | 4.08 | 0.07 | 2.097 |
| JSOSP+GCP [53] | 2.28 | 3.96 | 0.08 | 2.107 |
| KADI [54] | 2.07 | 5.16 | 0.08 | 2.437 |
| SSCBP [55] | 2.60 | 3.44 | 0.10 | 2.047 |
| ADCensus [56] | 2.42 | 4.10 | 0.09 | 2.203 |
| PM-Forest [57] | 1.32 | 1.91 | 0.15 | 1.127 |
| AdaptiveGF [58] | 2.44 | 5.71 | 0.17 | 2.773 |
| PatchMatch [59] | 2.47 | 2.99 | 0.21 | 1.890 |
| SemiGlobal [35] | 3.06 | 6.02 | 1.00 | 3.360 |
| StereoSONN [60] | 5.07 | 8.53 | 0.53 | 4.710 |
| Proposed | 2.2004 | 3.2498 | 0.6650 | 2.0384 |

Therefore, we suggest to refine each pair-wise SAD-Census cost volume separately, just as same as we refined our aggregated cost volume. This is shown in Fig.7. As we have already calculated individual cost volumes between the reference and each matching view (as in Fig.4), the only thing we require in this modified cost refinement is respective initial disparity results. We consider each *reference-matchview* pair as a separate stereo configuration, and apply the original stereo SGM algorithm to generate initial two-view disparity results. Next, we warp each of the matching views towards the reference to create binary masks, which then are used to identify cost indexes corresponding to erroneous matches that we need to increase. We add the same penalty that we used before, followed by the edge preserving filter.

Similarly, we repeat this cost refinement for each image pair for three iterations. Then we interpolate all the small baseline cost volumes to the size of the largest baseline volume according to (8), add them together as in (10), and aggregate using line refinement strategy in (9). This way, we create a more robust aggregated MBSC volume: $S(P,d)^{aggre}$, which is finally used to generate a more accurate MBS disparity result. The binary mask image, the warped matching view, and the refined disparity result after each recursive step for the Cones dataset is summarized in Fig.8(a), *(b)*, and *(c)*.

Additionally, we apply a few post processing techniques to fine-tune our final disparity results in the image space. We first apply a simple hole filling to identify occlusions that are identified after performing an LR-consistency check. For each pixel $P$ in $D(P)_L^{init}$ and its corresponding matching view point $Q$ in $D(P)_R^{init}$, we check whether the condition $dist|d_P - d_Q| \leq 1$. Iff it is satisfied, $P$ is chosen as a valid pixel, and invalidated as an outlier; otherwise. We fill these outliers

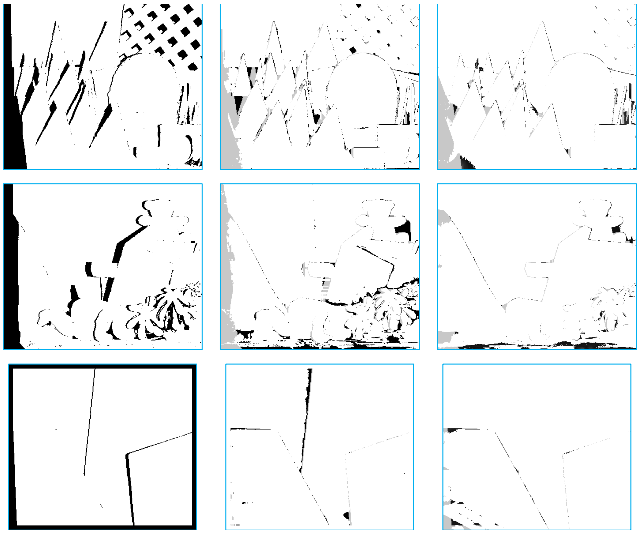


**FIGURE 9.** Non Occlusion mask comparison results for threshold 1.0. left: ground truth non occlusion mask, middle: original SGM non occlusion mask, right: our non occlusion mask.

by choosing the lowest disparity value of the spatially closest non-occluded pixels (inlier) that lie to the left and right on the same scanline. In order to remove the horizontal streaking effect after filling holes, we apply a simple weighted median filter.

Next, we apply a sub-pixel interpolation method as a quadratic polynomial is applied to increase the accuracy of disparity results. For each pixel $P$ in the disparity image, its interpolated disparity $d^{int}$ is calculated by considering two adjacent disparity values from left and right as in (16) shown at the bottom of this page, where $d_{before}$ and $d_{next}$ denote previous and next disparity indexes, respectively.

Finally, we apply a Weighted Least Squares filter in the form of a fast global smoother [47] on our disparity results to further refine them and to make them as much as closer to the ground truth images.

## V. EXPERIMENT RESULTS

In this section, we summarize a collection of experimental analyses we did to evaluate the accuracy of our proposed disparity mapping technique. We used 12 of the Middlebury datasets for our experiments, having the assumption of that images are properly rectified. We perform all these experiments offline using a general purpose 64 bit windows10 desktop with an Intel(R) Core(TM) i7-7700 CPU at 3.60GHz, and 16GB RAM.

For simplicity, we used the quarter-resolution images. Both qualitative and quantitative result comparisons are performed. The parameters we used when calculating SAD-Census volumes are summarized in Tab.1.

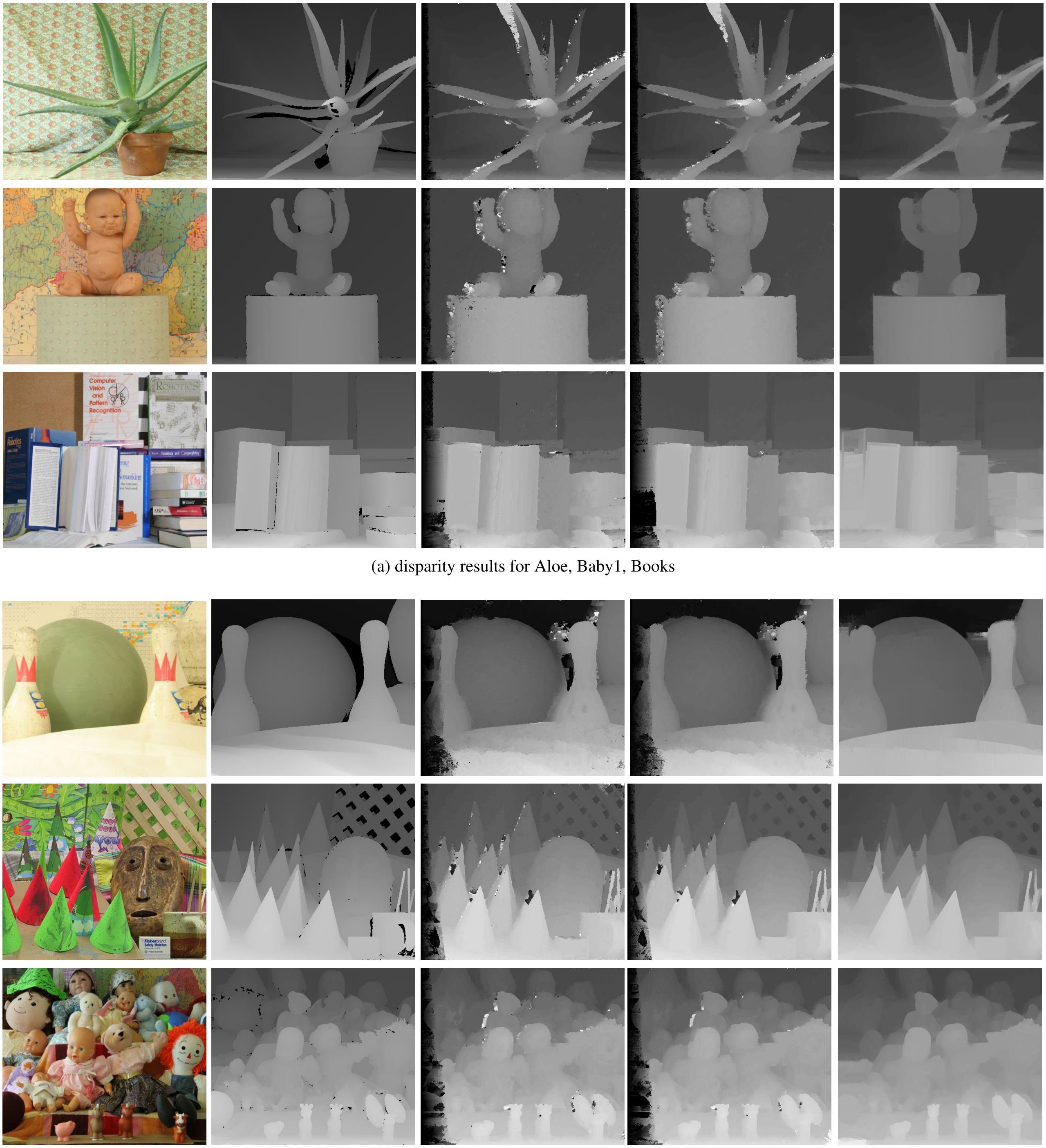
We performed the quantitative analyses by calculating the percentage of bad pixel error using the ground truth non occlusion masks for the Teddy, Cones and Venus datasets available in the Middlebury evaluation-$v2$ page. We additionally compared the results with few of the other state-of-the-art algorithms available in the evaluation page. For all these three datasets, we witnessed that the errors were all lower compared to the original SGM approach. However, one important fact worth noting is that we did not use the Tsukuba dataset for our evaluations. The reason was that all ground truth images including non occlusion mask available in the Middlebury evaluation site are given considering the middle image as the reference view out of five images. This violates the sequential increment constraint of baselines that we considered in our proposed method. The pixel errors for the Cones, Teddy, and Vensus datasets using the threshold value of 1.0 are summarized in Tab.2. Non occlusion masks between ours and original SGM are summarized in Fig.9.

The qualitative analyses are done comparing results between ours and the original MBS-SGM algorithm for all 12 datasets. These include Aloe, Baby1, Books, Bowling1, Cones, Dolls, Flowerpots, Lampshade1, Monopoly, Sawtooth, Teddy, and Venus, and are summarized in Fig.10 and Fig.11.

## VI. CONCLUSIONS

In this paper, we presented an iterative dense disparity estimating technique for a collection of multi-baseline stereo images with sequentially increasing baselines. The flow of the method consist of two individual steps: an initial multi-baseline matching cost calculation and disparity estimating step, and an iterative cost refinement step. In our first step, we employed a modified SAD-Census cost calculation method to compute all the pair-wise costs between the reference view and all the neighboring views. We interpolated the costs into the size of the cost associated with the largest disparity range and summed them to generate a multi-baseline matching cost volume. We exploited the scan line optimization method of the well-known SGM algorithm to aggregate this multi-baseline matching cost volume in sixteen directions and computed an initial disparity result using WTA approach. In the second step, we proposed to refine the aggregated matching cost volume by adapting a Gaussian modulating function and an edge preserving filter: rolling guidance-based guided filter. As we had already calculated pair-wise SAD-Census cost volumes between the reference and all its neighboring views in the first step, we suggested to compute individual stereo disparity maps between each view-pair, separately. Then we warped each matching view

$$d^{int} = d - [\frac{S(P, d_{next})^{aggre} - S(P, d_{prev})^{aggre}}{2(S(P, d_{next})^{aggre} - S(P, d_{prev})^{aggre} - 2S(P, d)^{aggre})}] \tag{16}$$

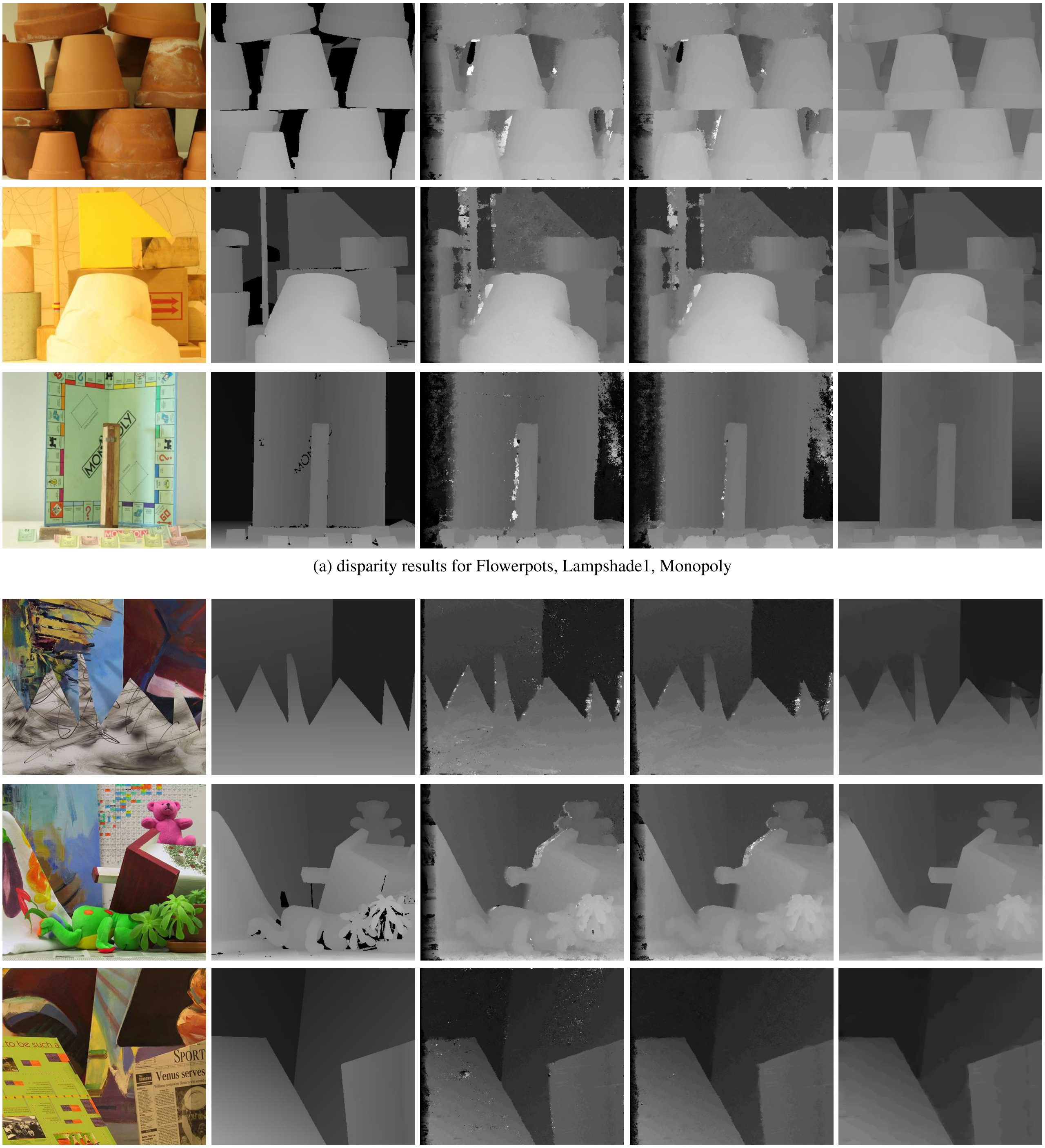


(a) disparity results for Aloe, Baby1, Books

(b) disparity results for Bowling1, Cones, Dolls

**FIGURE 10. Disparity results for the MBS Middlebury dataset. (left): reference, (left-middle): ground truth, (middle): disparity from original MBS-SGM, (right-middle): disparity from proposed method, (right): after applying WLS filter on result from proposed method.**

towards the view perspective of the reference to create binary mask images, which can be used to identify erroneously matched pixel locations. As white indexes depict incorrect correspondence matches, we added a high penalty value to the costs associated with these indexes in each pairwise cost volume to increase their values, such that their influence at disparity estimation would be minimized. Additionally, we applied the guided filter as a rolling guidance approach to further up-vote cost values as much as closer to ground truth data. We summed all the refined pair-wise costs again to generate a new multi-baseline SAD-Census cost volume and aggregated as in the first step. This we preformed recursively.



(a) disparity results for Flowerpots, Lampshade1, Monopoly

(b) disparity results for Sawtooth, Teddy, Venus

**FIGURE 11. Disparity results for the MBS Middlebury dataset. (left): reference, (left-middle): ground truth, (middle): disparity from original MBS-SGM, (right-middle): disparity from proposed method, (right): after applying WLS filter on result from proposed method.**

The final disparity map was filtered using an LR-consistency check, hole filling, sub-pixel interpolation, and WLS filter. We performed both qualitative and quantitative experiments to evaluate the accuracy of our proposed method. For this we used 12 Middlebury datasets. Through qualitative result analyses, we witnessed that our proposed method provided effective and efficient disparity results. Through quantitative analyses done for three of the datasets (Teddy, Cones, and Venus) available in the evaluation-*v2* page, we witnessed that our method showed a comparatively low average bad pixel value (2.0384 pixels) for the threshold of 1.0. As of future works, we are planning to extend our studies into image sets that are not pre-rectified, such as light field datasets, and to implement parallel programming for real-time or near real-time disparity estimations.

## REFERENCES

[1] R. Szeliski, *Computer Vision: Algorithms and Applications*. Berlin, Germany: Springer-Verlag, 2010.

[2] R. A. Hamzah and H. Ibrahim, "Literature survey on stereo vision disparity map algorithms," *J. Sensors*, vol. 2016, pp. 1–23, Dec. 2016.

[3] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[4] B. H. Bodkin, "Real-time mobile stereo vision," M.S. thesis, Dept. Elect. Eng. Comput. Sci., Univ. Tennessee, Knoxville, TN, USA, 2012.

[5] C. Stentoumis, L. Grammatikopoulos, I. Kalisperakis, and G. Karras, "On accurate dense stereo-matching using a local adaptive multi-cost approach," *ISPRS J. Photogramm. Remote Sens.*, vol. 91, pp. 29–49, May 2014.

[6] M. Valigi, S. Logozzo, and S. Affatato, "New challenges in tribology: Wear assessment using 3D optical scanners," *Materials*, vol. 10, no. 5, p. 548, May 2017.

[7] S. Seitz, "An overview of passive vision techniques," in *Proc. SIGGRAPH Course 3D Photography, Course Notes*, 1999, pp. 1–3.

[8] G. Bianco, A. Gallo, F. Bruno, and M. Muzzupappa, "A comparative analysis between active and passive techniques for underwater 3D reconstruction of close-range objects," *Sensors*, vol. 13, no. 8, pp. 11007–11031, Aug. 2013.

[9] O. Rahnama, T. Cavalleri, S. Golodetz, S. Waler, and P. Torr, "R3SGM: Real-time raster-respecting semi-global matching for power-constrained systems," in *Proc. Ternational Conf. Field-Program. Technol. (FPT)*, Dec. 2018, pp. 102–109.

[10] K. Zhang, Y. Fang, D. Min, L. Sun, S. Yang, and S. Yan, "Cross-scale cost aggregation for stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 5, pp. 965–976, May 2017.

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

[12] R. R. Orozco, C. Loscos, I. Martin, and A. Artusi, "HDR multiview image sequence generation: Toward 3d HDR video," in *High Dynamic Range Video*, New York, NY, USA: Academic, 2017, pp. 61–86.

[13] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.

[14] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," *Int. J. Comput. Vis.*, vol. 35, no. 3, pp. 269–293, 1999.

[15] Z. Zhang and J. Zou, "Image edge based efficient stereo matching," in *Proc. IEEE 8th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC)*, May 2019, pp. 180–185.

[16] Q. Dong and J. Feng, "Outlier detection and disparity refinement in stereo matching," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 380–390, Apr. 2019.

[17] H. Hirschmüller, P. R. Innocent, J. Garibaldi, "Real-time correlationbased stereo vision with reduced border errors," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 229–246, 2002.

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

[19] G. Egnal, "Mutual information as a stereo correspondence measure," Comput. Inf. Sci., Univ. of Pennsylvania, Philadelphia, PA, USA, Tech. Rep. MS-CIS-00-20, 2000.

[20] H. Hirschmuller, "Stereo vision based mapping and immediate virtual walkthroughs," Ph.D. dissertation, School Comput., De Montfort University, Leicester, U.K., 2003.

[21] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 492–504, Mar. 2009.

[22] C. Lei, J. Selzer, and Y.-H. Yang, "Region-tree based stereo using dynamic programming optimization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2378–2385.

[23] C. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 7, pp. 675–684, Jul. 2000.

[24] D. Kumari and K. Kaur, "A survey on stereo matching techniques for 3D vision in image processing," *Int. J. Eng. Manuf.*, vol. 6, no. 4, pp. 40–49, Jul. 2016.

[25] D. Honegger, T. Sattler, and M. Pollefeys, "Embedded real-time multi-baseline stereo," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 5245–5250.

[26] M. Poggi, F. Tosi, and S. Mattoccia, "Learning monocular depth estimation with unsupervised trinocular assumptions," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 324–333.

[27] S.-H. Baek and M. H. Kim, "Stereo fusion: Combining refractive and binocular disparity," *Comput. Vis. Image Understand.*, vol. 146, pp. 52–66, May 2016.

[28] Y. Furukawa and C. Hernández, "Multi-view stereo: A tutorial," *Found. Trends Comput. Graph. Vis.*, vol. 9, nos. 1–2, pp. 1–148, 2015.

[29] X. Hu and P. Mordohai, "Least commitment, viewpoint-based, multi-view stereo," in *Proc. 2nd Int. Conf. 3D Imag., Modeling, Process., Visualizat. Transmiss.*, Oct. 2012, pp. 531–538.

[30] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1362–1376.

[31] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 873–881.

[32] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm, "PatchMatch based joint view selection and depthmap estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1510–1517.

[33] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5483–5492.

[34] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR05)*, Feb. 2005, pp. 807–814.

[35] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[36] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 4, pp. 353–363, Apr. 1993.

[37] S. Bing Kang, R. Szeliski, and J. Chai, "Handling occlusions in dense multi-view stereo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, Dec. 2001, p. 1.

[38] T. Asai, M. Kanbara, and N. Yokoya, "3D modeling of outdoor environments by integrating omnidirectional range and color images," in *Proc. 5th Int. Conf. 3-D Digit. Imag. Model. (3DIM)*, Jun. 2005, pp. 447–454.

[39] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys, "Variable baseline/resolution stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[40] J. Li, H. Zhao, Z. Li, F. Gu, Z. Zhao, Y. Ma, and M. Fang, "A long baseline global stereo matching based upon short baseline estimation," *Meas. Sci. Technol.*, vol. 29, no. 5, May 2018, Art. no. 055201.

[41] N. Haala and M. Rothermel, "Dense multi-stereo matching for high quality digital elevation models," in *Photogrammetrie-Fernerkundung-Geoinformation*, vol. 2012, pp. 331–343, Aug. 2012.

[42] Z. Zhang, M. Gerke, G. Vosselman, and M. Y. Yang, "A patch-based method for the evaluation of dense image matching quality," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 70, pp. 25–34, Aug. 2018.

[43] D. Hernandez-Juarez, A. Chacón, A. Espinosa, D. Vázquez, J. C. Moure, and A. M. López, "Embedded real-time stereo estimation via semi-global matching on the GPU," *Procedia Comput. Sci.*, vol. 80, pp. 143–153, Jan. 2016.

[44] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. Eur. Conf. Comput. Vis.*, 1994, pp. 151–158.

[45] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[46] Q. Zhang, X. Shen, and J. Jia, "Rolling guidance filter," in *Proc. Eur. Conf. Comput. Vis.*, vol. 35, no. 6, 2014, pp. 815–830.

[47] D. Min, S. Choi, J. Lu, B. Ham, K. Sohn, and M. Do, "Fast global image smoothing based on weighted least squares," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5638–5653, Dec. 2014.

[48] J. Jiao, R. Wang, W. Wang, S. Dong, Z. Wang, and W. Gao, "Cost-volume filtering-based stereo matching with improved matching cost and secondary refinement," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 1–6.

[49] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 504–511, Feb. 2013.

[50] Q. Yang, "Hardware-efficient bilateral filtering for stereo matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 1026–1032, May 2014.

[51] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, "Bilateral filtering: Theory and applications," *Found. Trends Comput. Graph. Vis.*, vol. 4, no. 1, pp. 1–73, 2009.

[52] Y. Zhan, Y. Gu, K. Huang, C. Zhang, and K. Hu, "Accurate image-guided stereo matching with efficient matching cost and disparity refinement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1632–1645, Sep. 2016.

[53] J. Liu, C. Li, F. Mei, and Z. Wang, "3D entity-based stereo matching with ground control points and joint second-order smoothness prior," *Vis. Comput.*, vol. 31, no. 9, pp. 1253–1269, Sep. 2015.

[54] G. A. Kordelas, D. S. Alexiadis, P. Daras, and E. Izquierdo, "Enhanced disparity estimation in stereo images," *Image Vis. Comput.*, vol. 35, pp. 31–49, Mar. 2015.

[55] Y. Peng, G. Li, R. Wang, and W. Wang, "Stereo matching with space-constrained cost aggregation and segmentation-based disparity refinement," *Proc. SPIE*, vol. 9393, Mar. 2015, Art. no. 939309.

[56] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang, "On building an accurate stereo matching system on graphics hardware," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 467–474.

[57] X. Huang, C. Yuan, and J. Zhang, "A systematic stereo matching framework based on adaptive color transformation and patch-match forest," *J. Vis. Commun. Image Represent.*, to be published.

[58] Q. Yang, P. Ji, D. Li, S. Yao, and M. Zhang, "Near real-time stereo matching using adaptive guided filtering," *Image Vis. Comput.*, to be published.

[59] Q. Yang, P. Ji, D. Li, S. Yao, and M. Zhang, "PatchMatch stereo—Stereo matching with slanted support windows," *Bmvc*, vol. 11, pp. 1–11, Aug. 2011.

[60] M. Vanetti, I. Gallo, and E. Binaghi, "Dense two-frame stereo correspondence by self-organizing neural network," in *Proc. Int. Conf. Image Anal. Process.*, 2009, pp. 1035–1042.

**PATHUM RATHNAYAKA** received the B.Sc. and M.Sc. degrees in computer science and engineering from Kyungpook National University, Daegu, South Korea, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interests include stereo vision, stereo matching, and image processing.

**SOON-YONG PARK** received the B.Sc. and M.Sc. degrees in electronics engineering from Kyungpook National University, Daegu, South Korea, in 1991 and 1999, respectively, and the Ph.D. degree in electrical and computer engineering from the State University of New York at Stony Brook, in 2003. From 1993 to 1999, he was a Senior Research Staff with KAERI, South Korea. From 2011 to 2018, he was a Professor with the School of Computer Science and Engineering, Kyungpook National University, where he is currently working as a Professor with the School of Electronics and Engineering. His research interests include 3-D sensing and modeling, and multi-view 3D data processing.