Accurate Image-Guided Stereo Matching with Efficient Matching Cost and Disparity Refinement

Article *in* IEEE Transactions on Circuits and Systems for Video Technology · January 2015

DOI: 10.1109/TCSVT.2015.2473375

5 authors, including:

Zhan Yunlong
Chinese Academy of Sciences
9 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Keli Hu
Shaoxing University
14 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Accurate Image-Guided Stereo Matching With Efficient Matching Cost and Disparity Refinement

Yunlong Zhan, *Student Member, IEEE*, Yuzhang Gu, Kui Huang, Cheng Zhang, and Keli Hu

***Abstract*— Stereo matching is a challenging problem, and high-accuracy stereo matching is still required in various computer vision applications, e.g., 3-D scanning, autonomous navigation, and 3-D reconstruction. Therefore, we present a novel image-guided stereo matching algorithm, which employs the efficient combined matching cost and multistep disparity refinement, to improve the accuracy of existing local stereo matching algorithms. Different from all the other methods, we introduce a guidance image for the whole algorithm. This filter-based guidance image is generated by extracting the enhanced information from the raw stereo image. The combined matching cost consists of the novel double-RGB gradient, the improved lightweight census transform, and the image color. This cost measurement is robust against image noise and textureless regions in computing the matching cost. Furthermore, a new systemic multistep refinement process, which includes outlier classification, four-direction propagation, leftmost propagation, and an exponential step filter, is proposed to remove the outliers in the raw disparity map. Experiments on the Middlebury benchmark demonstrate our algorithm's superior performance that it ranks first among the 158 submitted algorithms. Moreover, the proposed method is also robust on the 30 Middlebury data sets and the real-world Karlsruhe Institute of Technology and Toyota Technological Institute benchmark.**

***Index Terms*— Double-RGB gradient, four-direction propagation, guidance image, lightweight census transform, stereo matching.**

I. Introduction

GENERATION of dense disparity map from a pair of stereo images is a popular topic in computer vision, because it plays a crucial role in many applications, including autonomous navigation, 3-D scanning, 3-D tracking, and 3-D reconstruction. The main problem is to estimate the correspondence between two pixels in each rectified image. Large numbers of methods have been proposed to solve this problem to achieve high accuracy and low computational cost. An extensive review of the stereo matching algorithms can be found in [1]. According to the evaluation strategy in [1], these stereo matching algorithms can be mainly classified into local and global methods. Global assumptions about energy or smoothness have been made in global methods to determine the disparity map. Local features (e.g., intensity levels) have been used in local methods to determine each pixel's disparity one by one. However, most of the existing global methods generate higher accuracy at the risk of execution time. Although local-feature-based methods enjoy fast running time, they require higher accuracy. Thus, this paper addresses the problem by improving the accuracy of local stereo matching.

As suggested in [1], most local stereo matching methods consist of four steps: 1) initial cost computation (computing the cost of each pixel); 2) cost (support) aggregation (aggregating the initial cost over the support region); 3) disparity computation (deciding each pixel's disparity level); and 4) disparity refinement (postprocessing to remove outliers). In addition to these steps, a preprocessing step is optional. In general, our algorithm follows the four suggested main steps and an appended preprocessing step. We give an introduction to each step before introducing our algorithm.

A. Related Works

1) Preprocessing: There are few specific studies on this issue. Bias–gain and histogram equalization [2], [3] were two early kinds. The background subtraction method in [4] was used to enhance the raw image for compensating radiometric differences. In many papers, a median filter is adopted as a default filter to suppress the noise in the raw image. Herein, we propose a filter-based algorithm to obtain enhanced content (i.e., guidance image) from the input image in this step.

The guidance image was early used in [5], in which a guided filter was proposed. By considering the content of the guidance image, the filtering output performed well in edge preserving. Thus, this filter was widely used in image processing, e.g., working as cost aggregation function in [6] and [7] to construct a constant time weighted median filter. However, the guidance image in those methods is the raw stereo image, so the effect is limited. In this paper, a novel filter-based guidance image is developed to preserve the image edges and smooth the homogeneous areas. Furthermore, this filter-based

Manuscript received January 17, 2015; revised April 9, 2015 and July 6, 2015; accepted August 19, 2015. This work was supported by the Strategic Priority Research Program through the Chinese Academy of Sciences under Grant XDA06020300. This paper was recommended by Associate Editor C. Zhang.

Y. Zhan, Y. Gu, and C. Zhang are with the Key Laboratory of Wireless Sensor Network and Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China (e-mail: qingqingzjin@gmail.com; gyz@mail.sim.ac.cn; mjcheng@mail.sim.ac.cn).

K. Huang is with the University of Science and Technology of China, Hefei 230026, China (e-mail: hkaustc@mail.ustc.edu.cn).

K. Hu is with the Department of Computer Science and Engineering, Shaoxing University, Shaoxing 312000, China (e-mail: ancimoon@gmail.com).

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/TCSVT.2015.2473375

guidance image, instead of the raw stereo image, provides the enhanced content for the whole stereo matching system, not just in cost aggregation function as in [6]. This image-guided stereo matching algorithm has not been studied before, since most papers did not consider the role of the guidance image in the algorithm, even not systemically proposing an image-guided stereo matching algorithm.

2) Matching Cost Measurement: Local stereo matching computes the cost of each pixel to get the initial cost volume. A wide spectrum of the initial cost measurements has been studied over the years. The common window-based initial cost measurements are the sum of absolute differences (SAD), squared differences, and normalized cross correlations (NCCs). However, those window-based methods were sensitive to noise, camera gain, and camera bias. The mean filter, Laplacian of Gaussian [8], [9], and the image gradient magnitude [10] had been taken to reduce the offset effeteness of the above methods. Unfortunately, the disparity maps of these filters were blurred. Nonparametric methods, such as mutual information (MI) algorithms [11], rank and census algorithms [12], wavelet-based algorithms [13], and local binary pattern [14], were proposed later as they were robust against outliers near object boundaries in window-based methods. The MI method [11] took the joint probability distribution function as the measurement of similarity. However, the kernel size of MI-based method was usually large. The contradiction is that the large kernel size leads to blurred object borders. The census transform method [12] encoded the local structure information into a string, and then a comparison of Hamming distance was taken to compute the similarity. This method is robust in both local and global stereo matching according to [15], and thus, there are many modifications on this measurement [16]–[18]. Nevertheless, the robustness of the census transform is kept with a relatively large kernel size (e.g., 7×7 in [16] and 9×7 in [19] and [20]). More recently, the traditional methods have been modified to get robust measurements in special conditions. Mahalanobis distance cross correlation [21] and adaptive NCC [22] were taken to get robust cost against the illumination variation in stereo images.

All of these methods have strengths and weaknesses, so some studies have been devoted to combining the strengths of multiple methods to achieve better performance. Klaus *et al.* [23] combined the intensity SAD and the image gradients for computing the initial matching cost. Mei *et al.* [19] combined the absolute differences (ADs) with the traditional census transform to achieve an impressive performance. These well-combined cost measurements achieve their goals and perform better than the single one, so we adopt this combined cost strategy in this paper.

Meanwhile, we find that the kernel size of the above-mentioned methods plays crucial roles in keeping robust. We are interested in improving the robustness of cost measurement and reducing the computational complexity, i.e., reducing the kernel size. Since the gradient image is insensitive to small noise, camera gain, and camera bias, we present a novel double-RGB gradient model to fully extract the gradient information from the raw stereo image and enhanced guidance image. Then, a novel combined matching cost measurement is proposed, which consists of the image color AD, the census transform, and the new double-RGB gradient AD. With the help of the robust information from the double-RGB gradient model, the window size of the census transform is sharply reduced without reducing the robustness of the cost measurement.

3) Cost Aggregation and Disparity Computation: The matching cost is usually aggregated by an aggregation function based on the assumption that the disparities in homogeneous areas have little difference and share the same disparity. Thus, a summation of the initial matching cost is computed for each pixel over the support region to remove the possible influence of noise. Most aggregation functions can be roughly classified into window-based method [24]–[26], filter-based method [6], [27], and segment-tree-based method [28], [29]. The exponential step method [26] is a special window-based method, which can aggregate cost in a large support region with low computational complexity. Later, we will further improve this exponential step function in our matching system. In our system, this exponential step function not only acts as aggregation function but also works as a novel postprocessing in the refinement step.

Almost all the papers take the winner-takes-all (WTA) strategy [1] to compute the raw disparity map. This strategy generates each pixel's disparity by choosing the disparity with the lowest aggregated cost value.

4) Disparity Refinement: To achieve higher accuracy, refinement is undertaken to identify and correct the outliers in the raw disparity map. Left–right consistency (LRC) checking [1] is the widely used method to detect the outliers. The detection method in [11] provided a way to distinguish the occluded outliers and the mismatched outliers. Several approaches, including region voting [19], vertical voting [20], cost spectrum peak analysis and removal [30], subpixel enhancement [19], [31], as well as filtering methods (e.g., using median filtering and weighted median filter [6], [7]), aimed at correcting the outliers in the raw disparity map. All the methods have flaws, and no single one method can remove all the outliers. Therefore, multistep refinement is adopted in [17], [19], and [31] to systematically remove outliers. This strategy achieved competitive results according to the literature. Thus, we propose a multistep refinement to identify and correct the outliers. Different from the above-mentioned methods, we classify the outliers into leftmost and inner kinds, and different approaches are taken to recover the outliers according to their kinds. Moreover, the image-guided exponential step filter further optimizes the disparity map.

B. Contributions

We survey the development of existing stereo approaches, and then propose an image-guided stereo matching system to improve the accuracy of local stereo matching. This system consists of four traditional main steps and an appended preprocessing step. First, a new filter-based guidance image is introduced in the preprocessing step. The guidance image will provide enhanced content for the whole system, which has not been studied before. Moreover, a novel double-RGB gradient

model is developed with the gradient content of both the guidance image and the raw image. This model will provide robust information for the cost measurement. Second, we propose a novel cost measurement that combines the information of the double-RGB gradient model, the lightweight census transform, and the image color together. This lightweight census transform is developed to reduce the computational amount for the first time. Third, an exponential step filter is adopted; this filter is not only used as an aggregation function, but also further improved as a novel refinement method in the postprocessing. Finally, the raw disparity map usually contains numerous outliers of different categories. We classify those outliers into inner ones and leftmost ones. A new systemic efficient refinement process, which includes four-direction propagation, leftmost propagation, and an exponential step filter, is then applied to recover the outliers according to their respective categories. Experiments on the Middlebury benchmark [32] and KITTI benchmark [33] demonstrate the effectiveness of our algorithm and show that our method is one of the state-of-the-art stereo matching algorithms.

The remainder of this paper is organized as follows. The filter-based guidance image model and the double-RGB model are introduced in Section II. In Section III, the combined cost measurement operator is proposed. An improved exponential step cost aggregation function is presented in Section IV. In Section V, a systematic multistep refinement process is adopted to achieve higher accuracy. The experimental process and results are elaborated in Section VI. A brief summary is given in Section VII.

II. Filter-Based Guidance Image Model and Double-RGB Gradient Model

A. Filter-Based Guidance Image Model

Traditionally, the raw image pairs are directly used in the stereo matching or simply with the median filter to suppress noise. However, the edges of the raw stereo image pairs may change slowly and many homogeneous areas are broken into smaller areas as a result of small noise or texture. In this paper, we propose a filter to improve these disadvantages. The filter not only preserves the strong boundaries and suppresses noise and weak boundaries to decrease their effect on the initial cost computation, but also smoothes the homogeneous areas to keep the disparity consistent. In other words, the filter enhances the strong boundaries and smoothes the homogeneous areas. This filter is called a mask filter; the filtered image, called a mask image, plays the role as the guidance image (GM) in the whole system.

The mask filter considers the spatiality and the color difference effect, which is similar to the bilateral filter used in [4] and the semblable bilateral filter used in [24]. Bilateral filter is rather effective for preserving edges and smoothing the homogeneous areas, which meets our requirements. However, in [24], it acted as a cost aggregation function to process the cost volume (3-D data) and was allocated a relatively large kernel size (e.g., 35×35), so its computation was expensive. In [4], a background-subtraction method was used to subtract the bilateral-filter-gained background to enhance the image for compensating radiometric differences. The enhanced image was used to calculate the matching cost with a limited improvement. In our algorithm, the mask filter is introduced to process the input image (only 2-D data) in the preprocessing step with a rather small half width R_{GM} (e.g., 5), and thus the computational amount is relatively small. In addition, the result is worked as a guidance image instead of being used to compute the matching cost directly. Guidance images that are generated by different filters are shown in Fig. 1.

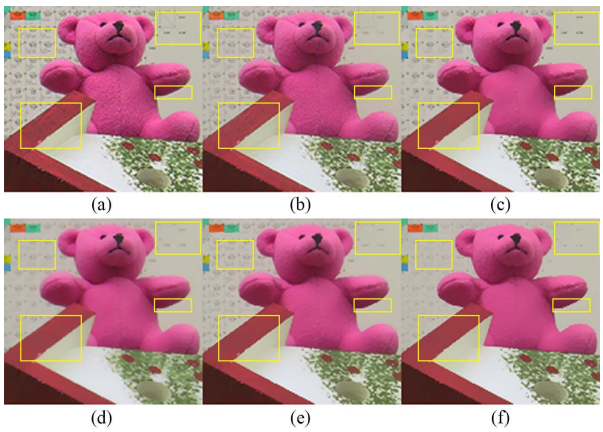

Fig. 1. Filtering results of the stereo image *Teddy* using different filters. (a) Raw image. (b) Using the guided filter [5]. (c) Using the domain transform filter [34]. (d) Using the WLS filter [35]. (e) Using the weighted median filter [7]. (f) Using the mask filter (i.e., the GM). Our filter well preserves the strong edges between different objects, and the tiny textures and noises in the homogeneous areas are clearly smoothed out as shown in the yellow rectangles, whereas the performances of the other filters are limited.

Now, we give a description of this filter-based guidance image model. Given a pixel $q(k, l)$ in the support window of pixel $p(i, j)$, with Δg_{pq} and Δc_{pq} representing the spatial Euclidian distance and the color difference in RGB space between p and q, respectively, and two constant parameters δ_d and δ_r used to adjust the spatial and color effect in the support window with kernel radius R, the filter weight of these two pixels is then

$$w(i, j, k, l) = \exp\left(-\Delta g_{pq}^2/\left(2\delta_d^2\right) - \Delta c_{pq}^2/\left(2\delta_r^2\right)\right). \quad (1)$$

According to the weight (1), local information is adequately used in the support window. The shorter the distance between the two pixels, the larger the weight will be. In addition, the closer the color of the two pixels, the larger the weight will be. The former helps the pixels to group by proximity, and the latter helps the pixels to group by similarity. The filtering output acts as the guidance image, and this filter helps the local pixels in the raw image to get a reasonable grouping, as shown in Fig. 1. Compared with the other filters, our mask filter achieves its goal and the effect is obviously in Fig. 1(f).

B. Double-RGB Gradient Model

Image gradient contains rich structural information (e.g., points and edges) and it is insensitive to illumination, so it is widely used in stereo matching algorithms [6], [10], [29]. Traditionally, gradient information is extracted from the intensity image, which is usually transformed from the RGB color space. Thus, some information is missing during

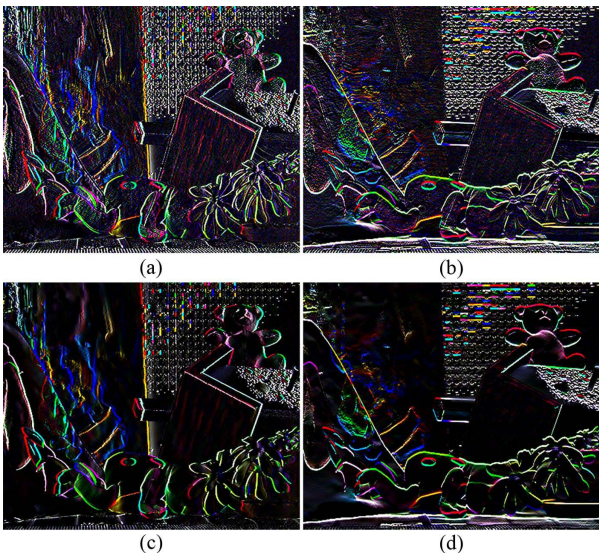

Fig. 2. Gradient results of the raw image *Teddy* and the corresponding GM. Gradient results of the raw image in (a) horizontal and (b) vertical directions. (c) and (d) GM results. (a) and (b) Local detailed information. (c) and (d) Strong global boundary information.

the conversion. In this work, gradient information is directly extracted from the RGB color space as it was done in [36]. But the difference is that we get the gradient information from both the raw stereo image and the guidance image.

Suppose that there is a pixel $p(x, y)$ with the RGB vector space (I_R, I_G, I_B). Each channel gradient of this color vector can be obtained according to

$$\nabla I(x, y) = (\nabla_x I, \nabla_y I) \tag{2}$$

where ∇_x and ∇_y are the partial derivatives, respectively, in the x and y directions. The GM and the raw stereo image are processed based on the gradient operator (2) to get two directional gradients. Then, by combining the information from both the raw image and the GM, the gradient model with double RGB vectors in one direction is generated as

$$\begin{aligned}\text{grad}_{\text{dir}} = \big(&w_1\nabla_{\text{dir}} I_R^{\text{raw}}, w_1\nabla_{\text{dir}} I_G^{\text{raw}}, w_1\nabla_{\text{dir}} I_B^{\text{raw}},\\ &w_2\nabla_{\text{dir}} I_R^{\text{GM}}, w_2\nabla_{\text{dir}} I_G^{\text{GM}}, w_2\nabla_{\text{dir}} I_B^{\text{GM}}\big)\end{aligned} \tag{3}$$

where the subscript dir indicates the direction along the x- or y-axis. The superscripts raw and GM refer to the gradient information of the raw image and the GM, respectively. w_1 and w_2 are allocated with empirical values to balance the effect of the raw image gradient and the guidance image gradient. As this new gradient vector combines both the raw image gradient and the guidance image gradient together, we call it double-RGB gradient.

As shown in Fig. 2, the raw image gradient contains abundant local details, but the GM gradient shows global main boundaries. The proposed double-RGB gradient model combines these two kinds of gradients to include the features of both local details and global main boundaries, so it is more robust than the traditional gradient. Thus, we adopt this model to measure the matching cost in the next section.

III. Combined Cost Measurement

As discussed in Section I-A, combined cost measurement achieves better performance than the single one. Thus, we adopt the combined cost strategy and propose a novel combined matching cost that consists of image color AD, lightweight census transform, and double-RGB gradient AD. The detailed implementation will be described in this section.

A. Measurement of the Double-RGB Gradient

Image gradient contains rich structural information, and it is insensitive to illumination; moreover, our double-RGB gradient is full of detailed local and global information, so we adopt it to measure the matching cost. The AD is chosen to measure the cost since it is widely used in cost measurement for its simple and easy implementation. Suppose that there is a pixel p and its corresponding pixel is q with disparity d. To suppress noise, we take the average AD of the reference image gradient vector grad^{ref} and the matching image gradient vector grad^{mat} according to

$$C_{\text{ADg,dir}}(p, d) = (1/6)\|\text{grad}_{\text{dir}}^{\text{ref}}(p) - \text{grad}_{\text{dir}}^{\text{mat}}(\text{q})\|_1 \tag{4}$$

to get directional gradient cost. The subscript dir indicates the direction along the x- or y-axis, and the directional gradient grad_{dir} can be gotten according to (3). The main computation of this measurement is to compute gradients in six channels and compute the Taxicab norm. This process can be easily implemented with low computational complexity.

B. Measurement of the Image Color

Measuring the AD of image color between two corresponding pixels is easy to be implemented; in addition, the image color might help to reduce the matching ambiguities in some repeating texture regions, and thus, calculating the AD of image color is a common cost measurement in stereo matching. In this study, we also chose this measurement and the cost is measured in color image directly. To suppress noise, the average AD of the reference image I^{ref} and matching image I^{mat} is taken in three RGB channels based on

$$C_{\text{ADc}}(p, d) = (1/3)\sum_{i\in\{R,G,B\}}\left|I_i^{\text{ref}}(p) - I_i^{\text{mat}}(q)\right|. \tag{5}$$

C. Measurement of the Lightweight Census Transform

The census transform [12] encodes the local structure information of the center pixel (other than the pixel intensity) over the support window and is robust in both local and global stereo matching [24]. For a pixel p with support window $W(m \times n)$, the census transform maps the neighboring pixels structure into a string vector (whose length is $m \times n$). Suppose that there are two pixels p and q and that q is in the support window of p (i.e., w_p). Then, the census transform of p is specified by

$$\text{cen}(p) = \otimes_{q\subseteq w_p} c(p, q) \tag{6}$$

where $\otimes$ denotes concatenation and $c(p, q)$ is a binary function. If the intensity value of pixel p is larger than that of q, $c(p, q)$ is equal to one; otherwise, $c(p, q)$ is zero. The census

transform cost of pixel p with level d is the Hamming distance of bit string $\text{cen}(p)$ and $\text{cen}(q)$

$$C_{\text{cen}}(p,d) = \text{cen}(p) \oplus \text{cen}(q) \tag{7}$$

where $\oplus$ denotes the Minkowski sum.

The computational complexity of the census transform cost measurement mainly depends on the window size M. However, the traditional window size is usually kept large (e.g., 7×7 in [16] and 9×7 in [19] and [20]) to be robust. In this algorithm, the window size M can be adjusted to a smaller one (e.g., 5×7) without reducing the cost measurement quality, because there is additional information from the double-RGB gradient cost measurement. This improvement reduces the computation for each pixel by factors of 4×7 and 2×7, respectively, thus sharply reducing computational amount. Therefore, it is referred to as lightweight census transform.

D. Combined Cost Measurement

The total initial cost C can be derived from the above three different costs: 1) the AD on image color; 2) the AD on new gradient model in both directions; and 3) the lightweight census transform of intensity image. This raw cost measurement fully considers the information from the GM and the raw stereo image. The cost measurement is

$$\begin{aligned} C(p,d) = {} & \alpha\rho(C_{\text{ADc}}, \lambda_{\text{ADc}}) + \beta\rho(C_{\text{cen}}, \lambda_{\text{cen}}) \\ & + \xi\rho(C_{\text{ADgx}}, \lambda_{\text{ADgx}}) + \varepsilon\rho(C_{\text{ADgy}}, \lambda_{\text{ADgy}}) \end{aligned} \tag{8}$$

which employs the truncation function

$$\rho(C,\lambda) = \min(C,\lambda) \tag{9}$$

where λ is the threshold to ensure the cost not severely biased by one measurement. Here, α, β, ξ, and ε are the weight parameters used to control the influence of each cost measurement. In contrast to other methods, in our method, two directional gradient costs are combined with different weights to obtain the final gradient cost, since these two directional costs are not equally important. The subsequent experimental results verify this weight strategy.

IV. Cost Aggregation and Disparity Computation

The initial cost measurement (8) is usually full of noise, which influences the disparity map. Exponential step aggregation function runs fast with adaptive weight and provides a good tradeoff between accuracy and speed according to [26] and [37]. Thus, we adopt this aggregation function and we further add a new average factor N in the aggregation function. In our algorithm, this function not only acts as an aggregation function here but also works as a refinement filter in postprocessing. Besides, the disparity computation is utilized to compute the raw disparity map. Both operators are introduced in detail in the following.

A. Cost Aggregation Function

Cost aggregation is the most important step to reduce noise; however, it usually consumes most of the time in stereo matching system. The purpose of our algorithm is to reduce its computational complexity and to keep it robust. The exponential step aggregation function meets this requirement. This function was proposed in [26] and developed in [37]. We adopt this aggregation structure, but we further improve the weight strategy. Fig. 3 shows the structure in horizontal direction (the structure in vertical direction is similar). Three iterations are performed in Fig. 3 with step sizes (i.e., s) equal to 1 (range of $[-1, 1]$), then 3 (range of $[-4, 4]$), and then 9 (range of $[-13, 13]$). Each computation involves 3 pixels, and only 9 (3×3) pixels are referred to in three iterations with the computation ranging from -13 to 13, which greatly diminishes the computation compared with the conventional algorithm. The step size could be slightly tuned as shown in [37], and the number of iterations could also be tuned according to different applications in practice.

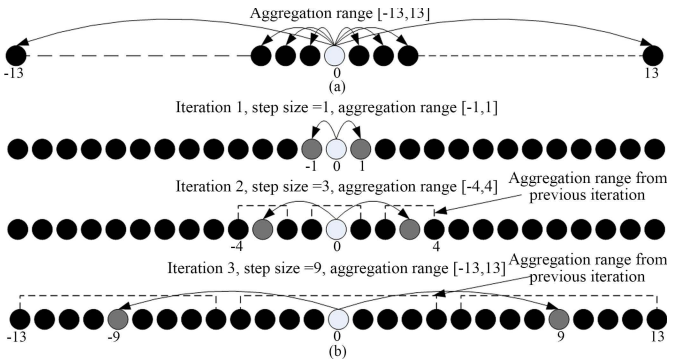

Fig. 3. Example of 1-D cost aggregation. (a) Traditional cost aggregation method to aggregate cost in the range $[-13, 13]$. (b) Exponential step aggregation method to aggregate cost in three simple iterations in the range $[-13, 13]$ (after the structure in [26]).

The most significant aspect of this function is that the cost is separately aggregated in horizontal and vertical directions instead of directly on the $N \times M$ fixed window. Different interactions can be applied to achieve different ranges in two directions. The original aggregation function directly aggregated the cost of two weighted neighbors and itself. We improve the function by adding an average factor N, which is different from what was done in [37]. Hence, the cost is updated in each iteration with only three involved pixels based on

$$\begin{aligned} C_{\text{agg}}(p,d) = {} & C(p,d) + (w(p,p+s)C(p+s,d) \\ & + w(p,p-s)C(p-s,d))/N \end{aligned} \tag{10}$$

where N is equal to the number of existing endpoints, which averages the contribution of two endpoints. This average factor improves the aggregation performance by reducing about 0.61% average error percentage in the *disc* region on four standard Middlebury data sets. $w(p,q)$ is the weight of pixels p and target pixel q as

$$w(p,d) = \exp(-\Delta g_{pq}/\lambda_d - \Delta c_{pq}/\lambda_I) \tag{11}$$

which considers the intensity difference and spatial distance effects, similar to the mask filter weight (2). But here $I(.)$ is the proposed GM, which provides enhanced content for the aggregation function.

B. Disparity Computation

We take the WTA strategy [1] to compute the disparity map from the aggregated cost volume C_{agg} in this step. WTA strategy is the most widely used method to determine the disparity for each pixel. Suppose that $d_{\max}$ and $d_{\min}$ represent the maximum and minimum disparity, respectively. This strategy generates each pixel's optimal disparity d_p^* by choosing the disparity with the lowest aggregated cost value in all allowed disparities according to

$$d_p^* = \arg \min_{d \in [d_{\min}, d_{\max}]} C_{\text{agg}}(p, d). \tag{12}$$

We take this strategy (12) to compute the left raw disparity map D_L and the right raw disparity map D_R. Until now, the raw disparity maps are generated.

V. Multistep Disparity Refinement

The raw disparity map usually contains lots of outliers, especially near depth discontinuities and in occlusion regions, but it is difficult to remove all of them with only one method. We instead handle these outliers through a novel systematic multistep process. Outliers are first classified into two types, and then corrected by the following approaches, including four-direction propagation, leftmost propagation, exponential step filtering, and median filtering, according to their types.

A. Outlier Detection and Classification

In the beginning stage, an LRC check [1] is used to detect the outliers in D_L. According to this check, the feasible corresponding pairs are those found both in the reference image and in the matching image, so the disparity value of the conjugate pairs in the left disparity map D_L and right disparity map D_R should be equal. Suppose pixel $p(x, y)$ is in D_L, then the corresponding pixel in D_R should be $q(x - \max(D_L(x, y), 0), y)$, where $\max(D_L(x, y), 0)$ takes the larger value between $D_L(x, y)$ and 0. If the pixel's disparity value fails in the LRC check according to

$$D_L(x, y) = D_R(x - \max(D_L(x, y), 0), y) \tag{13}$$

it is an outlier, and its disparity is incorrect.

The outliers can be further divided into leftmost outliers, inner outliers, and rightmost outliers according to their positions in the image. The leftmost outliers in D_L come from the right image information lost in computing cost measurement, while the inner outliers and rightmost outliers come from occlusion or mismatch. Thus, both the inner and rightmost outliers are labeled as inner outliers without distinguishing. Different strategies are established to correct them later.

B. Simplified Cross-Skeleton Support Region

We take the propagation strategy to correct the outliers, whereupon the inner outliers should be recovered with exactly reliable neighboring disparities. Before taking processes to recover the outliers, we first compute the support region for each outlier, and then the following processes depend on this structure. The support region constructed in [38] is adaptive and competitive, hence we adopt this structure for our algorithm. But we simplify the support region for each pixel to contain only two directional regions [i.e., $H(p) \cup V(p)$]. Furthermore, we compute only the support region for outliers instead of all the pixels as is done in [38] so as to reduce the computational complexity. The simplified cross skeleton is shown in Fig. 4. Suppose pixel q is in one of pixel p's directions, thus an adaptive linear threshold to control the direction expansion in Fig. 4 is as

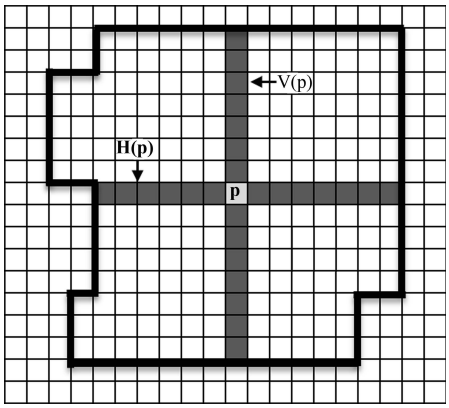

Fig. 4. Simplified cross-skeleton support region $H(p) \cup V(p)$ for pixel p. The cross skeleton of pixel p consists of a horizontal direction $H(p)$ and a vertical direction $V(p)$. Each direction can expand in two arms, e.g., the horizontal direction consisting of left and right arms.

$$D_s(L_{pq}) = D_{\max}(L_{\max} - L_{pq})/L_{\max} \tag{14}$$

where D_s and L_{pq} are the color dissimilarity threshold and the spatial Euclidian distance between p and q, respectively. $D_{\max}$ and $L_{\max}$ are two constant parameters to control the color and spatial effect, respectively. If the color Euclidian distance D_{pq} is below the threshold D_s and the spatial Euclidian distance L_{pq} is shorter than the largest spatial distance $L_{\max}$, the support region of p expands to q until q does not meet the conditions. Each arm expands its support region in this way.

C. Two-Step Four-Direction Propagation

Four-direction propagation is proposed to obtain valid information from four directions to correct the inner outliers, while the traditional two-direction propagation method (e.g., scanline propagation [20]) obtains only the valid information from left and right. Our method can obtain more valid information. In this paper, this novel four-direction propagation employs two steps to recover the inner outliers. First, the outlier seeks reliable information from its support area by searching in its support area along four arms to find four reliable disparities, i.e., *dl*, *dr*, *du*, and *dd*. Function *min(dl, dr)*, which takes the smaller value between *dl* and *dr*, is defined as *dlr*. Similarly, function *min(du, dd)* is defined as *dud*. Then, the outlier disparity is updated based on

$$d_p^* = \begin{cases} null, & \text{if } dlr \text{ and } dud \text{ not existing} \\ dlr, & \text{else if only } dlr \text{ existing} \\ dud, & \text{else if only } dud \text{ existing} \\ (dlr + dud)/2, & \text{else if } |\, dlr\text{-}dud \,| \le 2 \\ null, & \text{otherwise} \end{cases} \tag{15}$$

and this step runs for two iterations to fully fill the outliers.

After the first step, the left outliers are almost the occlusion outliers, which can be filled with the background disparities. Thus, the above process is improved by removing the searching area limitation. The same search is applied along each outlier's four arms to find four reliable disparities without the support area limitation. Then, the disparity is updated depending on

$$d_p^* = \begin{cases} null, & \text{if } dlr \text{ and } dud \text{ not existing} \\ dlr, & \text{else if only } dlr \text{ existing} \\ dud, & \text{else if only } dud \text{ existing} \\ min(dlr, dud), & \text{otherwise.} \end{cases} \tag{16}$$

This step also runs for two iterations until all the inner outliers are filled with reliable disparities.

D. Leftmost Propagation

The leftmost propagation is designed to recover the leftmost outliers, since they are special outliers whose right image does not have related information in matching cost computation. Most local methods recover them by propagating the right neighbors' disparity level, without considering the variation of the circumjacent disparities. We propose a method involving both of them.

The variation trend is acquired by analyzing the reliable pixel disparity from the first reliable pixel in the outliers' right arm to the next reliable pixel until the disparity jumps more than one disparity value or the pixel color difference exceeds a threshold. Then, the trend is classified into three types: increasing, decreasing, or remaining the same. The leftmost outliers are then corrected, from right-part to left-part, with the disparity value from the inner reliable pixels following the above-mentioned trend. One segmentation (i.e., the outlier left arm's length) is propagated each time. The recovered disparity is interpolated with an increment, a decrement, or the same value as the right reliable disparity according to the trend. This step runs for two iterations to fill all the leftmost outliers.

E. Exponential Step Filter

This step is proposed to further remove the error disparities that are not detected by the LRC check or that are recovered with error values through the above-mentioned approaches, especially the disparities around the discontinuous areas. A simple filter used in the cost aggregation meets this goal, because this filter has a good tradeoff between accuracy and speed; moreover, it can make use of the enhanced content in the guidance image (GM). We call this filter the ES filter.

The exponential step (ES) filter works as follows. For each pixel p with disparity d, its new cost is computed based on

$$C_{\text{new}}(p, d) = \min(\mu d_{\max}, |d - \widehat{D_L}(p)|) \tag{17}$$

where $\widehat{D_L}(p)$ is the refined disparity map with all the above steps. Then, the ES filter is taken to aggregate this cost with the guidance image, and the parameters are the same as the cost aggregation process. Next, the WTA strategy, according to (12), is used to compute the new disparity map D_L^*. In practice, we run this step for two iterations to fully suppress the noise. However, it is better not to use this filter more than three times; otherwise, some boundaries will be oversegmented. In addition, morphology methods can be further used to improve the performance, e.g., removal of small areas. It is an optional method according to different applications. Finally, a 3×3 median filter is applied to smooth the disparity map.

The effectiveness of each refinement process is shown in Fig. 7. Detailed arguments of this systemic refinement process are presented in the next section.

TABLE I
PARAMETER SETTINGS FOR THE PROPOSED ALGORITHM

PRM	Value	PRM	Value	PRM	Value
R_{GM}	5	L_{max}	31	μ	0.2
W_{cen}	5×7	D_{max}	24	ε	0.489
λ_I	15/255	α	0.244	λ_{ADc}	18/255
λ_d	15/255	β	0.116	λ_{cen}	21/255
w_1/w_2	1/2	ξ	0.151	λ_{ADg}	8/255

VI. EXPERIMENTS AND DISCUSSION

In this section, experiments are conducted to study the performance of this novel matching algorithm. We mainly focus on the following objectives: 1) the performance of the proposed algorithm compared with other methods; 2) the performance of the guidance image (GM); 3) the performance of the double-RGB gradient model; 4) the performance of the combined cost measurement; 5) the performance of the multistep refinement; and 6) the adaptability of the proposed algorithm.

A. Experiment Environment and Evaluation Methods

In this section, we mainly consider two benchmarks. The indoor scene is the Middlebury benchmark [32], and the outdoor scene is the KITTI benchmark [33]. Experiments will be mainly carried out on the online Middlebury benchmark, which is built to evaluate the performance of stereo matching algorithms and with which one can upload disparity maps to reproduce the results. This test benchmark includes four image pairs (i.e., *Teddy*, *Cones*, *Venus*, and *Tsukuba*) and the corresponding ideal disparity maps. The performance of each disparity map is evaluated with the percentage of absolute disparity error in three different regions, which are the nonoccluded area (i.e., *nonocc*), the whole image area (i.e., *all*), and the area around discontinuities (i.e., *disc*). Experiments will also be carried out on KITTI benchmark to test our algorithm's adaptability. The parameters of our algorithm are listed in Table I. In order to get compelling experimental results, all the parameters are kept constant for the following experiments (except for the new *method*^{c} in Table II). We set $L_{\max}$ and $D_{\max}$ as in [38]. λ_I and λ_d are the same as in [37].

B. Performance of the Proposed Algorithm

Experiments are carried out on the standard Middlebury benchmark [32] to test the performance of the proposed method. The disparity results are shown in Fig. 5, and the evaluation results are shown in Table II. The right subscript

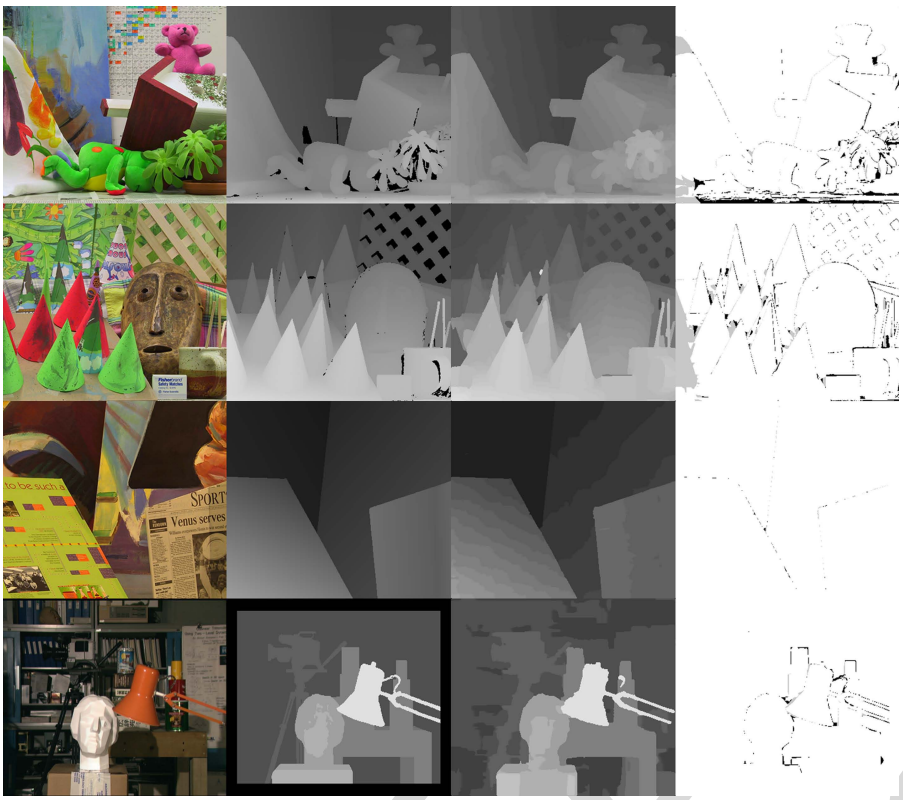

Fig. 5. Our results of the Middlebury data set. From top to bottom: *Teddy*, *Cones*, *Venus*, and *Tsukuba*. From left to right: left stereo image, ground truth, our disparity map, and the error of our disparity map. The disparity map errors in nonoccluded and occluded regions are marked in black and gray, respectively.

TABLE II
EVALUATION RESULTS OF ERROR PERCENTAGES ON MIDDLEBURY DATA SET (ERROR THRESHOLD = 1.0)

Algorithm		Our method	LCU[a]	TSGO[b]	JSOSP + GCP [39]	Our method[c]	AD-Census [19]	AdaptingBP [23]
Tsukuba	nonocc	0.93_{10}	1.06_{18}	0.87_{4}	0.74_{1}	1.06_{17}	1.07_{21}	1.11_{24}
	all	1.37_{12}	1.34_{8}	1.13_{1}	1.34_{9}	1.46_{18}	1.48_{19}	1.37_{11}
	disc	5.05_{12}	5.50_{16}	4.66_{6}	3.98_{1}	5.73_{23}	5.73_{24}	5.79_{26}
Venus	nonocc	0.07_{3}	0.07_{2}	0.11_{9}	0.08_{4}	$\mathbf{0.07_{1}}$	0.09_{5}	0.10_{7}
	all	0.17_{4}	0.26_{17}	0.24_{11}	0.16_{1}	0.20_{7}	0.25_{14}	0.21_{10}
	disc	1.04_{3}	1.03_{2}	1.47_{12}	1.15_{4}	$\mathbf{1.02_{1}}$	1.15_{4}	1.44_{11}
Teddy	nonocc	4.08_{19}	3.68_{15}	5.61_{42}	3.96_{17}	4.09_{19}	4.10_{20}	4.22_{22}
	all	5.98_{8}	9.95_{36}	8.09_{19}	10.1_{37}	5.9_{18}	6.22_{9}	7.06_{17}
	disc	11.4_{20}	10.4_{14}	13.8_{35}	11.8_{21}	11.3_{20}	10.9_{16}	11.8_{22}
Cones	nonocc	2.14_{9}	1.63_{2}	1.67_{3}	2.28_{19}	2.37_{22}	2.42_{25}	2.48_{29}
	all	6.97_{14}	6.87_{12}	6.16_{2}	7.91_{33}	7.14_{17}	7.25_{19}	7.92_{35}
	disc	6.27_{8}	4.82_{2}	4.95_{3}	6.74_{22}	6.97_{26}	6.95_{26}	7.32_{34}
Average Error		**3.79**	3.89	4.06	4.18	3.94	3.97	4.23
Average Ranking		$\mathbf{10.2_{1}}$	12.0_{2}	12.2_{3}	14.1_{4}	14.9_{4}	16.8_{5}	20.7_{6}

[a] From the anonymous paper 'Using local cues to improve dense stereo matching,' which was submitted to *CVPR* 2015.
[b] From the paper 'Accurate stereo matching by two step global optimization', which was proposed by M. Mozerov and J. van Weijer and submitted to *IEEE Trans. Image Process.* in 2014.
[c] Tuning the parameters α, β, ξ, and ε in Table I to get the results. All the evaluation results are percentages and come from the Middlebury benchmark [32]. The right subscript of each error percentage is the ranking in each region.

of each error percentage is the ranking in each region. As demonstrated in Table II, our algorithm ranks first among the 158 submitted algorithms in the Middlebury evaluation. The results rank better than the *ADCensus* method [19], which once ranked first and was the best local stereo matching algorithm. Our algorithm performs well on all the stereo pairs in three regions. Moreover, our method's average error is 3.79%, which is the lowest value among all the methods in Table II. The competitive performance of our algorithm achieves our aim of developing an accurate local stereo matching system. Besides, our algorithm has the potential to achieve better results in a certain region by tuning the

TABLE III
EVALUATION RESULTS OF DIFFERENT GUIDANCE IMAGES ON MIDDLEBURY DATA SET (ERROR THRESHOLD = 1.0)

Algorithm		Our GM	Raw-image -based GM	MedFilter[a] -based GM	[5]-based GM	[7]-based GM	[34]-based GM	[35]-based GM
Tsukuba	nonocc	0.93_{10}	1.12_{25}	0.96_{11}	0.98_{14}	0.94_{10}	$\mathbf{0.89_{9}}$	0.92_{10}
	all	1.37_{12}	1.65_{32}	1.49_{19}	1.52_{19}	1.40_{14}	$\mathbf{1.35_{11}}$	1.43_{18}
	disc	5.05_{12}	5.52_{16}	4.97_{11}	5.31_{15}	5.10_{14}	4.83_{11}	$\mathbf{4.65_{6}}$
Venus	nonocc	$\mathbf{0.07_{3}}$	0.22_{50}	0.10_{7}	0.12_{10}	0.10_{6}	0.12_{10}	0.11_{7}
	all	$\mathbf{0.17_{4}}$	0.42_{46}	0.25_{14}	0.31_{22}	0.19_{5}	0.28_{18}	0.28_{19}
	disc	$\mathbf{1.04_{3}}$	2.83_{61}	1.39_{9}	1.51_{13}	1.42_{10}	1.57_{16}	1.43_{10}
Teddy	nonocc	$\mathbf{4.08_{19}}$	4.90_{26}	4.42_{22}	4.46_{22}	4.23_{22}	4.30_{22}	4.09_{19}
	all	$\mathbf{5.98_{8}}$	7.17_{17}	6.77_{14}	6.57_{13}	$\mathbf{5.98_{8}}$	6.04_{8}	6.21_{8}
	disc	$\mathbf{11.4_{20}}$	13.7_{34}	12.4_{25}	12.6_{26}	11.8_{21}	11.9_{22}	11.5_{20}
Cones	nonocc	2.14_{9}	2.11_{7}	2.13_{8}	$\mathbf{2.07_{5}}$	2.16_{11}	2.27_{16}	2.17_{11}
	all	6.97_{14}	7.51_{22}	$\mathbf{6.90_{14}}$	7.09_{16}	7.25_{18}	7.33_{19}	7.06_{16}
	disc	6.27_{8}	6.19_{8}	6.23_{8}	$\mathbf{6.05_{5}}$	6.34_{9}	6.67_{18}	6.36_{10}
Average Error		**3.79**	4.45	4.00	4.05	3.91	3.96	3.85
Average Ranking		$\mathbf{10.2_{1}}$	28.7_{16}	13.5_{3}	15.0_{4}	12.3_{3}	15.0_{4}	12.8_{3}

[a] Median-filter-based guidance image. All the evaluation results are percentages and come from the Middlebury benchmark [32]. The right subscript of each error percentage is the ranking in each region.

parameters α, β, ξ, and ε in Table I. The results of *nonocc* and *disc* regions on *Venus* can achieve the first ranking among all the algorithms in the new *method*[c], but this method sacrifices the performance of other images. Thus, we can adopt different parameters according to diverse applications.

C. Performance of the Guidance Image

The guidance image (GM) has been widely used in the stereo matching system, including cost computation, cost aggregation, and multistep refinement. Now, we study the GM performance in our algorithm. Experiments are performed with our proposed GM, the raw stereo image, and the median filter generated image. Raw-image-based GM takes the raw stereo images as the guidance images (i.e., no guidance image). Median-filter-based GM employs a median filter (kernel size: 3×3) as mask filter to process the raw stereo image in each channel to get the GM. The experimental results are listed in Table III, and we can observe that the proposed GM plays an important role in improving the average error. Except for the *nonocc* and *disc* regions on *Cones*, the proposed GM obviously improves the performance in all the other regions compared with the results of raw-image-based GM. The average ranking increases from 28.7 up to 10.2, and the average error is reduced from 4.45% down to 3.79%. Moreover, the proposed GM performs better than the median-filter-based GM. The median-filter-based GM performs better than the raw-image-based GM, and the average error is 4.00%, which is relatively low and ranks well. All of the excellent performances indicate that the proposed image-guided matching system is valid and effective. What is more, the experimental results confirm the importance of the GM in the stereo matching system and that our mask filter is more robust than the median filter. The reason for the limited improvement on *Cones* is that some textures of the raw images are very thin and permeated by background near the boundaries. Thus, the mask filter fails to work well with a relatively small kernel size ($R_{\mathrm{GM}} = 5$).

In addition, we further consider the performances of more different guidance images in our system. Four widely used filters, including the guided filter [5], the weighted median filter [6], [7], the domain transform filter [34], and the weighted least squares (WLS) filter [35], are taken to get the corresponding guidance images. Experiments are carried out with the MATLAB codes provided in [5], [7], [34], and [35]. According to the codes, we employ the guided filter and the WLS filter as mask filters to process the raw images on each channel separately, while the weighted median filter and domain transform filter work as mask filters to process the raw images directly. In addition, we tune all the filters' parameters to acquire relatively good results, which are also shown in Table III. According to Table III, all the filter-based guidance images have positive impact on the quality improvement compared with raw-image-based GM. However, a different filter has different effects on the stereo images. Guided filter performs well on *Cones*, and the errors in *nonocc* and *disc* regions are relatively small. The weighted median filter has a significant influence in *all* regions on *Teddy* and *Venus* compared with the other filters. The domain transform filter and the WLS filer play a clear effect on *Tsukuba*. Moreover, all the filter-based guidance images perform well with a relatively high ranking compared with the raw-image-based GM. This experiment demonstrates that the proposed system has good compatibility to adopt different guidance images. Furthermore, more studies could be carried out to find a more robust guidance image.

D. Performance of the Double-RGB Gradient and the Combined Cost Measurement

We first discuss the performance of the double-RGB gradient model. Two traditional gradients [i.e., the gray-image-based gradient (GrayG) and the color-image-based gradients (ColorG)] are taken for comparative study. Our proposed stereo matching system is Image guided

TABLE IV
PERFORMANCE OF DIFFERENT COST MEASUREMENTS

Algorithm	Avg. Error	Avg. Ranking	Algorithm	Avg. Error	Avg. Ranking
IGSM	3.79	10.2_{1}	GDxy	4.87	33.5_{16}
ColorG	4.02	17.8_{5}	[6]	4.77	34.8_{19}
GrayG	3.89	15.4_{4}	[29]	4.77	$34,8_{19}$
AD	5.10	50.0_{43}	[20]	3.93	15.5_{4}
Census	4.14	32.1_{16}	[19]	4.53	39.5_{22}

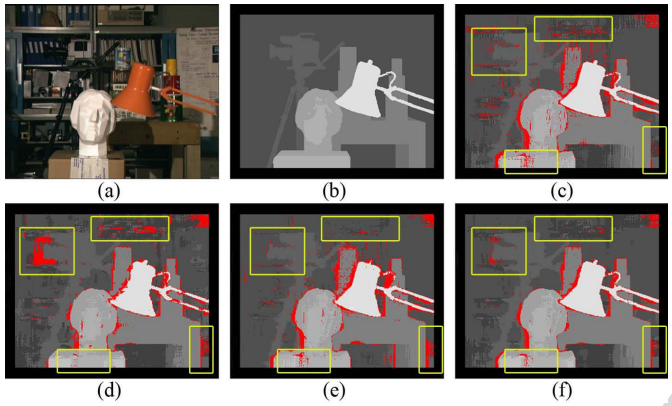

Fig. 6. Disparity maps computed by different cost measurements. (a) Left stereo image. (b) Ground truth. (c) Operator: the AD of image color. (d) Operator: the lightweight census transform. (e) Operator: the AD of the double-RGB gradient. (f) Operator: the proposed measurement. The incorrect pixels are marked in red.

stereo matching (IGSM). Experiments are carried out on the standard Middlebury data sets to verify the effectiveness of each gradient. Traditional gradients are tested in the proposed combined cost measurement by replacing the double-RGB gradient operator. The experimental results are shown in Table IV. It is demonstrated in Table IV that our proposed double-RGB gradient model is competitive in both accuracy and ranking compared with two traditional gradient-based combined cost measurement. In other words, the double-RGB gradient is more effective.

Our proposed cost measurement consists of the AD of image color (AD), the lightweight census transform (Census), and the AD of the double-RGB gradient (GDxy). Experiments are carried out to study the performance of each operator. As shown in Fig. 6, the incorrect pixels are marked in red and several yellow rectangles are drawn to show the accumulations of incorrect pixels in different disparity maps. There are some drawbacks for each individual operator: the census transform fails to process the repetitive local structures; the large textureless regions are hard for the pixel-based AD operator; and the proposed double-RGB gradient cannot well handle the slowly changing or tiny boundaries. However, the combined cost measurement successfully reduces some incorrect pixels caused by individual measurements, which can be clearly observed in Fig. 6(f) that the incorrect pixels are reduced in those red rectangles. Moreover, discrete mismatches are also well suppressed. The left mismatches are mainly around the boundaries or in the occlusion regions. According to the experiments, the combined cost measurement absorbs the advantages of each operator, because the pixel-based AD mainly improves the performance in *disc* region; the census transform handles well in the *all* region; and the gradient-based AD reduces the mismatches in *nonocc* region.

In quantitative comparison, the proposed operator successfully reduces the average error percentage compared with three individual cost measurement-based measurements (i.e., AD, Census, and GDxy) based on Table IV. Besides, some existing combined cost measurements are also taken into consideration: AD of image color and AD of gradient on both x and y directions were used in [6] and [29]; AD of image color with traditional census transform was used in [19] and [20] (truncation functions in them are different). These combined cost measurements are set with the papers the parameters recommended by [6], [29], [19] and [20]. As can be observed in Table IV, our proposed combined cost measurement has advantages over these existing combined cost measurements when used in our algorithm system.

E. Performance of the Multistep Refinement

We will verify the effectiveness of the multistep refinement in this section. As shown in Fig. 7, the performance of each process is measured by the average error in three different regions. Fig. 7(a)–(c) shows the detailed performance on each image data set, and the average results can be observed in Fig. 7(d). This multistep postprocessing successfully reduces the error percentage in different regions: for *all* region, the errors are mainly corrected by four-direction propagation, leftmost propagation, and ES filter, and meanwhile, four-direction propagation massively reduces the error percentage; for *nonocc* region, four-direction propagation is the most effective way to handle outliers. ES filter performs different on four images. At the same time, the effect of leftmost propagation is limited since the leftmost regions contain few *nonocc* regions; for *disc* region, the error percentage is significantly reduced by four-direction propagation; furthermore, leftmost propagation and ES filter further reduce the error percentage. In quantitative comparison, the average error percentages of the four raw disparity maps are reduced by 6.47% in the *all* region, 2.39% in *nonocc* region, and 9.39% in *disc* region after taking this multistep postprocessing.

Now, we check the performance of each refinement step. The first step is outlier detection. However, there is an error percentage increment after taking the outlier detection as shown in Fig. 7. This phenomenon is caused by the fact that there are some reliable pixels marked as outliers owing to the existence of noise, but the error percentage quickly decreases with the following processes. As shown in Fig. 7(d), four-direction propagation massively reduces the average error percentage in *all*, *nonocc*, and *disc* regions by 12.47%, 10.41%, and 20.45%, respectively. These improvements show that two-step four-direction propagation is appropriate for filling inner outliers. Leftmost propagation is effective in the *all* and *disc* regions (e.g., *all* region on *Teddy* with a 4.9% decrement and *disc* region on *Venus* with a 2.3% decrement), but not in *nonocc* region, because the leftmost regions mainly exist in the leftmost outliers. The significant effects are clearly shown in Fig. 7(a) and (c), especially on *Venus* and *Teddy*.

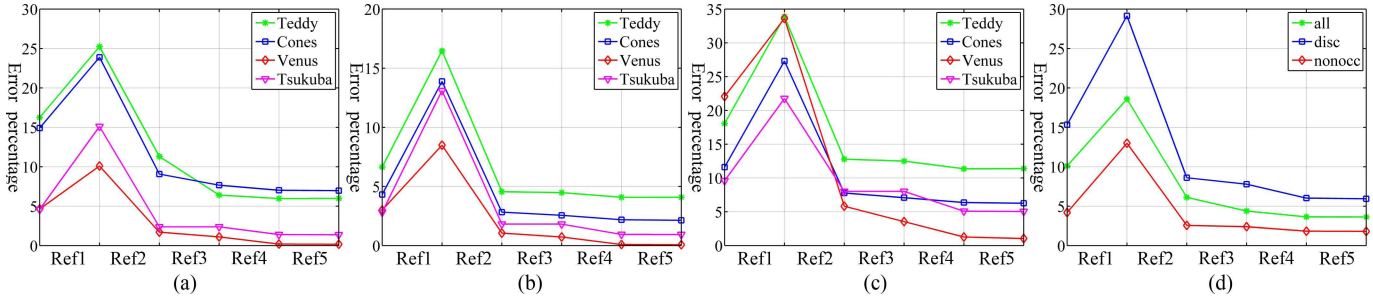

Fig. 7. Average error percentages from raw disparity map to the refined disparity map. (a) Error percentages in the *all* region. (b) Error percentages in the *nonocc* region. (c) Error percentages in the *disc* region. (d) Average error percentages of three regions. The refinement operators from *Ref1* to *Ref5* are outlier detection, four-direction propagation, leftmost propagation, ES filter, and median filter.

ES filter further reduces the error percentage in all the three regions, especially on *Venus* and *Tsukuba*, since this filter fully reprocesses the disparity maps. Finally, a simple median filter is applied to smooth the disparity map. The detailed performances of each refinement step in three regions are, respectively, shown in Fig. 7(a)–(d). In summary, significant improvements are obtained for each raw disparity map. Furthermore, a systematic refinement (with the above-mentioned refinement steps) guarantees an effective postprocessing.

F. Adaptability of the Proposed Algorithm

Many existing algorithms verify only the performance on the standard Middlebury data set [32]. However, we will carry out new experiments on other data sets to test the proposed algorithm's performance for more reliable evaluation. New data sets in Middlebury [32] and the outdoor KITTI street data set [33] are taken into consideration.

The Middlebury benchmark [32] contains many data sets involving more complex scenes. We use Middlebury 2001 (7 stereo pairs), Middlebury 2003 (2 stereo pairs), Middlebury 2005 (6 stereo pairs), and Middlebury 2006 (15 stereo pairs) data sets. Thus, we have 30 stereo pairs in total and we denote it by $M30$. Moreover, four state-of-the-art algorithms [6], [29], [40], [41] are included to make a comparison. Table V shows the quantitative evaluation results on $M30$ in *nonocc* region with one pixel error threshold, and some disparity maps of the new data sets are shown in Fig. 8. The results of the state-of-the-art algorithms are reproduced using the codes provided in [6], [29], [40], [41]. Our proposed method is denoted as IGSM, which produces an competitive performance with 20 most accurate results on $M30$ in Table V. Moreover, IGSM outperforms all the other methods with the highest average ranking 2.07 and the lowest average error percentage 4.99%.

Finally, we test our method IGSM and the above-mentioned four methods on KITTI data set [33]. All the images in this data set are captured under real-world condition, thus many image pairs contain large textureless regions, e.g., sky, road, and walls, and variable brightness conditions, e.g., shades of trees. This data set contains 195 test image pairs and 194 training image pairs with corresponding ground-truth disparity maps for evaluating stereo matching methods. However, these ground-truth disparity maps should be filled in with background to generate dense disparity maps. As shown in Fig. 9, disparity maps that are generated by our method are more smooth and contain less noise, especially fewer large black areas, compared with all the other methods, so our method performs well in KITTI benchmark. Both of the above experiments demonstrate the good adaptability of our method.

TABLE V
ERROR PERCENTAGES IN NONOCCLUSION REGION

Algorithm	IGSM	[29]	[40]	[41]	[6]
Teddy	**4.08**	6.79	6.15	7.86	6.99
Cones	**2.14**	2.75	2.60	3.24	3.23
Venus	**0.07**	0.28	0.20	0.74	1.09
Tsukuba	**0.93**	1.86	1.60	2.43	2.30
Barn1	**0.20**	0.58	0.41	0.69	0.85
Barn2	**0.34**	0.47	0.38	0.75	1.06
Bull	**0.10**	0.35	0.13	0.59	0.59
Poster	**0.12**	0.62	0.39	1.34	1.54
Sawtooth	**0.64**	0.84	0.70	1.24	1.25
Art	**7.83**	9.45	9.79	10.07	9.41
Books	10.84	10.92	11.06	**9.99**	11.32
Dolls	**4.75**	5.65	6.24	5.49	6.06
Laundry	15.35	15.11	16.14	14.41	**12.74**
Moebius	9.94	10.09	10.38	**8.62**	9.01
Reindeer	**5.07**	5.96	7.96	6.54	7.41
Aloe	**4.59**	6.76	4.95	5.09	6.93
Baby1	**3.45**	4.23	5.18	4.72	4.58
Baby2	**3.77**	5.68	5.99	14.43	4.90
Baby3	**3.89**	6.97	7.57	5.96	5.09
Bowling1	19.24	14.24	14.51	18.77	**11.90**
Bowling2	7.46	8.77	10.12	10.50	**7.15**
Cloth1	0.50	1.23	0.42	**0.38**	1.73
Cloth2	2.97	4.32	**2.71**	2.93	4.52
Cloth3	**1.24**	2.28	1.66	1.65	2.67
Cloth4	**1.02**	2.16	1.36	1.25	2.16
Flowerpots	17.09	15.57	17.87	15.04	**11.53**
Lampshade1	**7.41**	11.97	12.76	12.20	10.30
Lampshade2	**6.86**	22.77	23.98	21.37	15.16
Rocks1	5.04	4.42	4.47	**3.71**	5.14
Rocks2	2.67	2.42	**2.29**	2.37	2.54
Average Error	**4.99**	6.18	6.33	6.48	5.70
Average Ranking	**2.07**	3.37	2.40	3.27	3.90

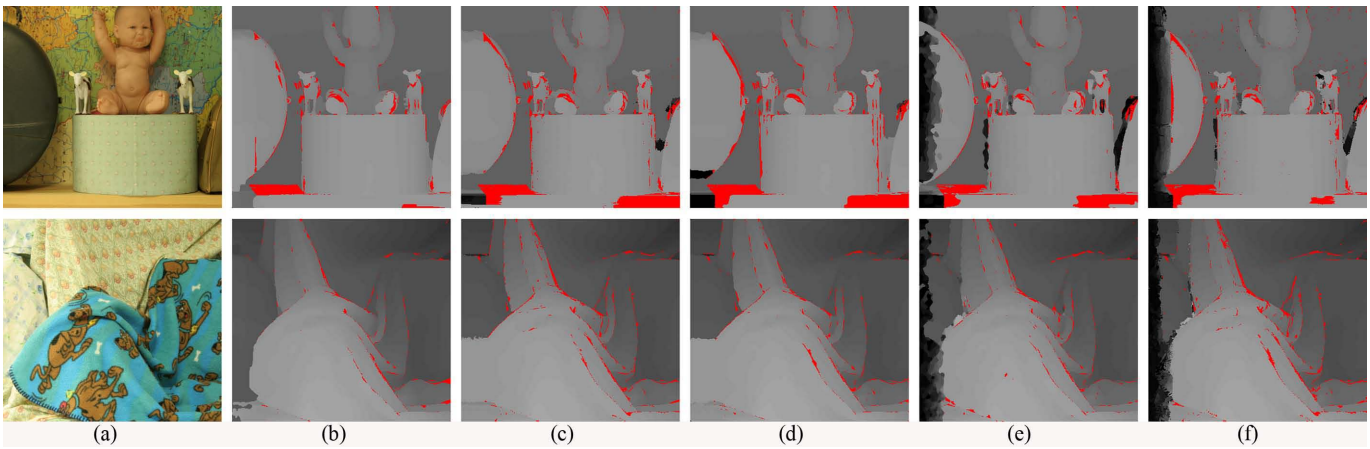

Fig. 8. Disparity maps of *Baby3* and *Cloth3*. Incorrect pixels in *nonocc* regions are marked in red. (a) Left images. (b) Our proposed method results. (c) CostFilter [6] results. (d) SSMP [41] results. (e) SegmentTree [29] results. (f) CrossScale [40] results.

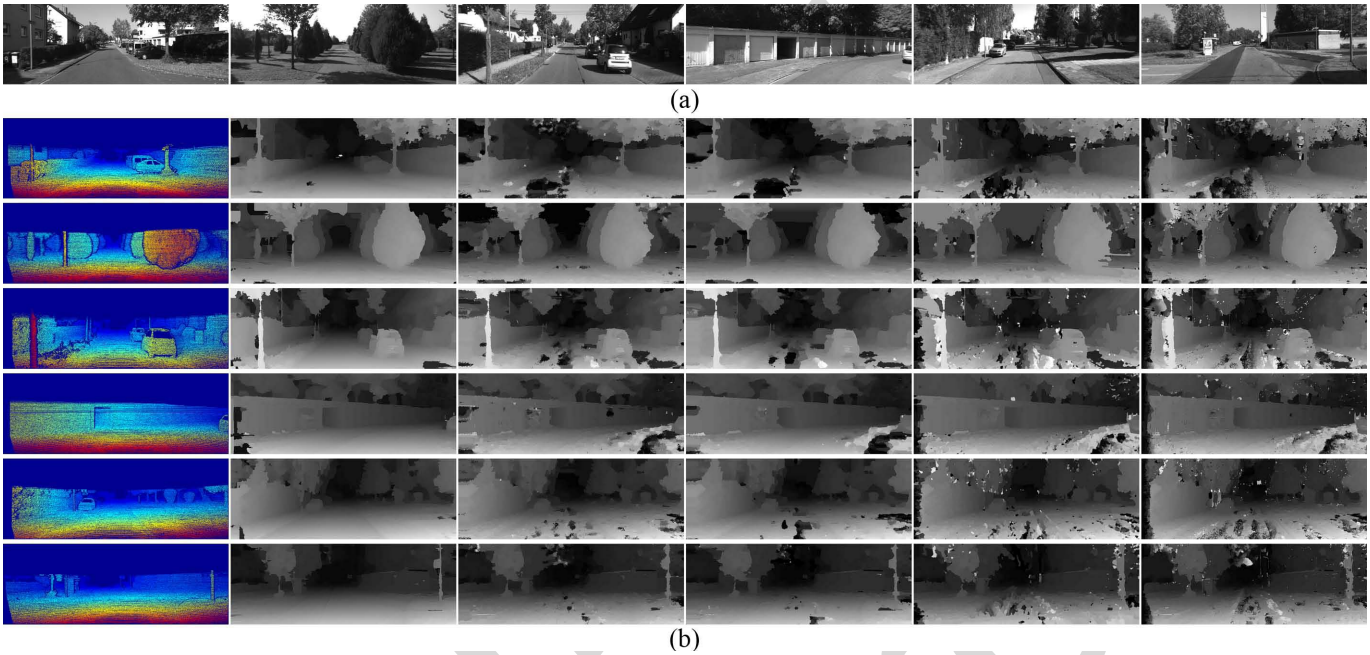

Fig. 9. Disparity maps on KITTI data set. (a) Left images (frames #000008, #000009, #000017, #000023, #000024, and #000050). (b) Disparity maps of different methods. From left to right: our proposed method, CostFilter [6], SSMP [41], SegmentTree [29], and CrossScale [40].

Based on the above experimental results, we conclude that our image-guided matching system is valid and effective. Moreover, it is obvious that, to ensure that our stereo matching system works well, the guidance image model, the double-RGB gradient model, the combined cost measurement, and the multistep refinement are all required. Hence, to keep the stereo matching system effective, it is quite important to develop a suitable guidance image. Unlike the guidance image, appropriate parameters are important for both the double-RGB gradient model and the combined cost measurement. Besides, refinement is supposed to be taken according to the qualities of outliers. These topics open up additional future researches on guidance image generation, parameter adjustment, and outlier classification.

In addition, we briefly analyze the complexity of the proposed algorithm. Suppose that the image size is $W * H$, the disparity range is D, and the support window width is N. The complexity of each step is characterized as $O(B_0WH)$ for the guidance image, $O(kB_1WH)$ for the double-RGB gradient, $O(kB_1WHD + B_2WHD + B_3WHD)$ for combined cost measurement, $O(WHD)$ for disparity computation, and $O(B_4WHD)$ for cost aggregation and disparity refinement. In those expressions, B_0 [complexity is $O(N^2)$] is the support window size of the mask filter; k is the number of gradient directions according to (3); B_1 and B_2 are the color channels of the double-RGB gradient and the raw input image, respectively; B_3 [complexity is $O(N^2)$] is the support window size of the lightweight census transform; and $O(B_4)$ [$O(B_4) = O(\log N)$ according to [26]] is the complexity to aggregate N pixels in the cost aggregation and the ES filter. In order to reduce the complexity of our algorithm, the size of B_0 and B_3 should be decreased. According to the parameters

in Table I, we can find that the majority of the computational complexity comes from the combined cost measurement, and the census transform is most time consuming. However, our proposed lightweight census transform sharply reduces the window size B_3 compared with the traditional ones, and thus, our proposed algorithm can cut down the computational amount in some way.

VII. Conclusion

A high-accuracy local stereo matching system has been proposed in this paper. The image-guided matching structure is valid and can be extensively adopted. Our system is efficient based on these key factors: the filter-based guidance image, the double-RGB gradient, the combined cost measurement, the exponential step aggregation structure, and the systematic efficient multistep refinement, including outlier classification, four-direction propagation, leftmost propagation, and exponential step filtering. Experiments demonstrate that our algorithm offers excellent high accuracy performance in both indoor and outdoor environments. This algorithm can be regarded as one of the state-of-the-art stereo methods. Furthermore, this study can be extended in terms of optimizing the computational complexity and reducing the number of parameters. Future research will be applied to additional indoor and outdoor scenes.

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, pp. 7–42, 2002.

[2] M. A. Gennert, "Brightness-based stereo matching," in *Proc. IEEE 2nd Int. Conf. ICCV*, Dec. 1988, pp. 139–143.

[3] I. J. Cox, S. Roy, and S. L. Hingorani, "Dynamic histogram warping of image pairs for constant image brightness," in *Proc. IEEE Int. Conf. ICIP*, vol. 2. Oct. 1995, pp. 366–369.

[4] A. Ansar, A. Castano, and L. Matthies, "Enhanced real-time stereo using bilateral filtering," in *Proc. 2nd Int. Symp. 3D Data Process., Visualizat., Transmiss.*, Sep. 2004, pp. 455–462.

[5] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. Int. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14. [Online]. Available: http://research.microsoft.com/en-us/um/people/kahe/eccv10/index.html

[6] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3017–3024. [Online]. Available: https://www.ims.tuwien.ac.at/publications/tuw-202088

[7] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu, "Constant time weighted median filtering for stereo matching and beyond," in *Proc. IEEE Int. Conf. ICCV*, Dec. 2013, pp. 49–56. [Online]. Available: http://research.microsoft.com/en-us/um/people/kahe/

[8] K. Konolige, "Small vision systems: Hardware and implementation," in *Robotics Research*. London, U.K.: Springer-Verlag, 1998, pp. 203–212.

[9] H. Hirschmüller, P. R. Innocent, and J. Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 229–246, 2002.

[10] E. P. Baltsavias and D. Stallmann, "Spot stereo matching for digital terrain model generation," in *Proc. 2nd Swiss Symp. Pattern Recognit. Comput. Vis.*, 1993, pp. 61–72.

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

[12] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. 3rd Int. Eur. Conf. Comput. Vis.*, 1994, pp. 151–158.

[13] I. Sarkar and M. Bansal, "A wavelet-based multiresolution approach to solve the stereo correspondence problem using mutual information," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 4, pp. 1009–1014, Aug. 2007.

[14] V. D. Nguyen, D. D. Nguyen, S. J. Lee, and J. W. Jeon, "Local density encoding for robust stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 12, pp. 2049–2062, Dec. 2014.

[15] H. Hirschmüller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1582–1599, Sep. 2009.

[16] W. S. Fife and J. K. Archibald, "Improved census transforms for resource-optimized stereo vision," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 1, pp. 60–73, Jan. 2013.

[17] J. Jiao, R. Wang, W. Wang, S. Dong, Z. Wang, and W. Gao, "Local stereo matching with improved matching cost and disparity refinement," *IEEE Multimedia*, vol. 21, no. 4, pp. 16–27, Oct./Dec. 2014.

[18] S.-C. Pei and Y.-Y. Wang, "Color invariant census transform for stereo matching algorithm," in *Proc. 17th IEEE Int. Symp. Consum. Electron. (ISCE)*, Jun. 2013, pp. 209–210.

[19] 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. Workshop*, Nov. 2011, pp. 467–474.

[20] X. Sun, X. Mei, S. Jiao, M. Zhou, and H. Wang, "Stereo matching with reliable disparity propagation," in *Proc. IEEE Int. Conf. 3DIMPVT*, May 2011, pp. 132–139.

[21] S. Kim, B. Ham, B. Kim, and K. Sohn, "Mahalanobis distance cross-correlation for illumination-invariant stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 11, pp. 1844–1859, Nov. 2014.

[22] Y. S. Heo, K. M. Lee, and S. U. Lee, "Robust stereo matching using adaptive normalized cross-correlation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 807–822, Apr. 2011.

[23] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proc. 18th Int. Conf. ICPR*, vol. 3. 2006, pp. 15–18.

[24] 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, Apr. 2006.

[25] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 7, pp. 1073–1079, Jul. 2009.

[26] W. Yu, T. Chen, and J. C. Hoe, "Real time stereo vision using exponential step cost aggregation on GPU," in *Proc. 16th IEEE ICIP*, Nov. 2009, pp. 4281–4284.

[27] C. C. Pham and J. W. Jeon, "Domain transformation-based efficient cost aggregation for local stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 7, pp. 1119–1130, Jul. 2013.

[28] Q. Yang, "A non-local cost aggregation method for stereo matching," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 1402–1409.

[29] X. Mei, X. Sun, W. Dong, H. Wang, and X. Zhang, "Segment-tree based cost aggregation for stereo matching," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 313–320. [Online]. Available: http://www.cs.albany.edu/~xmei/resource/page/segment-tree.html

[30] Q. Yang, P. Ji, D. Li, S. Yao, and M. Zhang, "Fast stereo matching using adaptive guided filtering," *Image Vis. Comput.*, vol. 32, no. 3, pp. 202–211, 2014.

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

[32] D. Scharstein and R. Szeliski. (2010). *Middlebury Stereo Evaluation*. [Online]. Available: http://vision.middlebury.edu/stereo/eval/

[33] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 3354–3361. [Online]. Available: http://www.cvlibs.net/datasets/kitti/

[34] E. S. L. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," *ACM Trans. Graph.*, vol. 30, no. 4, 2011, Art. ID 69. [Online]. Available: http://inf.ufrgs.br/~eslgastal/DomainTransform/

[35] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski, "Edge-preserving decompositions for multi-scale tone and detail manipulation," *ACM Trans. Graph.*, vol. 27, no. 3, 2008, Art. ID 67. [Online]. Available: http://www.cs.huji.ac.il/~danix/epd/

[36] Y. Geng, Y. Zhao, and H. Chen, "Stereo matching based on adaptive support-weight approach in RGB vector space," *Appl. Opt.*, vol. 51, no. 16, pp. 3538–3545, 2012.

[37] J. Zhang, J.-F. Nezan, M. Pelcat, and J.-G. Cousin, "Real-time GPU-based local stereo matching method," in *Proc. IEEE Conf. DASIP*, Oct. 2013, pp. 209–214.

[38] C. Stentoumis, L. Grammatikopoulos, I. Kalisperakis, E. Petsa, and G. Karras, "A local adaptive approach for dense stereo matching in architectural scene reconstruction," in *Proc. 5th Int. Workshop 3D-ARCH*, Feb. 2013, pp. 1–8.

[39] 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, 2015.

[40] K. Zhang *et al.*, "Cross-scale cost aggregation for stereo matching," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 1590–1597. [Online]. Available: https://github.com/rookiepig/CrossScaleStereo#cross-scale-cost-aggregation-for-stereo-matching-cvpr-2014

[41] B. Ham, D. Min, C. Oh, M. N. Do, and K. Sohn, "Probability-based rendering for view synthesis," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 870–884, Feb. 2014. [Online]. Available: http://www.di.ens.fr/~bham/pbr/index.html

Yunlong Zhan (S'14) received the B.S. degree in electronics engineering from Shanghai Jiao Tong University, Shanghai, China, in 2011. He is currently working toward the Ph.D. degree in information and communication engineering with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai.

His research interests include stereo vision, pattern recognition, and image processing.

Yuzhang Gu received the Ph.D. degree from Tokyo Institute of Technology, Tokyo, Japan.

He is an Associate Professor with Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China. His research interests include computer vision, target tracking, object detection, and 3-D video.

Kui Huang received the B.S. degree in computer science and technology from Soochow University, Suzhou, China, in 2012. She is currently working toward the master's degree in computer science and technology with University of Science and Technology of China, Hefei, China.

Her research interests include computer vision, image parallel processing, and GPU parallel acceleration.

Cheng Zhang is currently working toward the Ph.D. degree with Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China.

His research interests include computer vision, pattern recognition, and 2-D and 3-D face recognition.

Keli Hu received the B.S. degree in communication engineering from Hangzhou Dianzi University, Hangzhou, China, in 2009, and the Ph.D. degree in information and communication engineering from Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China, in 2014.

He has been a Teacher with the Department of Computer Science and Engineering, Shaoxing University, Shaoxing, China, since 2011. His research interests include artificial intelligence, pattern recognition, computer vision, and image processing.