Chang et al. EURASIP Journal on Advances in Signal Processing (2017) 2017:27
DOI 10.1186/s13634-017-0462-3

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Robust stereo matching with trinary cross color census and triple image-based refinements

Ting-An Chang, Xiao Lu and Jar-Ferr Yang*

Abstract

For future 3D TV broadcasting systems and navigation applications, it is necessary to have accurate stereo matching which could precisely estimate depth map from two distanced cameras. In this paper, we first suggest a trinary cross color (TCC) census transform, which can help to achieve accurate disparity raw matching cost with low computational cost. The two-pass cost aggregation (TPCA) is formed to compute the aggregation cost, then the disparity map can be obtained by a range winner-take-all (RWTA) process and a white hole filling procedure. To further enhance the accuracy performance, a range left-right checking (RLRC) method is proposed to classify the results as correct, mismatched, or occluded pixels. Then, the image-based refinements for the mismatched and occluded pixels are proposed to refine the classified errors. Finally, the image-based cross voting and a median filter are employed to complete the fine depth estimation. Experimental results show that the proposed semi-global stereo matching system achieves considerably accurate disparity maps with reasonable computation cost.

Keywords: TCC census transform, Cost aggregation, Range winner-take-all, Image-based refinements

1 Introduction

The measure of the distance of the scene for robotic systems [1, 2], self-directed vehicles [3], or 3D video broadcasting systems [4, 5] is an important research topic in computer vision. For 3D video broadcasting, a small number of selected views, which include the color texture frames and gray depth maps, are coded by the 3D-HEVC coders [6, 7]. In the receivers, the 3D TV set decodes all texture frames and depth maps with the 3D-HEVC decoder and use a depth image-based rendering (DIBR) system to generate more virtual views for naked-eye multi-view 3D displays [8, 9]. In case that the users possess the naked-eye multi-view 3D displays, the side-by-side or top-and-bottom stereo packing formats should further involve not only real-time stereo matching to estimate the depth information but also these displays which also need the depth image-based rendering (DIBR) process to produce the multi-view synthesized videos. Due to the high computation of stereo matching, a simple and accurate stereo matching algorithm is needed for multi-view 3D displays. Physically, the depth map could be measured by various sensors, such as laser or infrared radar by using the concept of time-of-flight to obtain accurate depth information but with disadvantages of low resolution and high cost. With multiple cameras [10, 11], the stereo vision technologies [12–14] to extract the depth information become a low-price and high-resolution approach. With horizontally placed cameras, the distance estimation of each pixel, called stereo matching, searches the best correspondence of the same scene point in two different viewing images [15, 16]. The horizontal displacement of the paired pixels in two viewing images is called the disparity. If the parameters of capturing cameras are known, the disparity map can be easily transformed to distance (depth) information.

Stereo matching, which is an active research topic in computer vision, could estimate a dense disparity map from a pair of images if their inherent ambiguities can be properly resolved. How to accurately estimate the disparity map under different scene conditions, such as smooth regions, discontinuities, and occluded areas, is

* Correspondence: jarferryang@gmail.com
Department of Electrical Engineering, Institute of Computer and Communication Engineering, National Cheng Kung University, Tainan, Taiwan

the most difficult problem. A survey of stereo matching was conducted by Scharstein and Szeliski [17]. Two well-known global stereo matching approaches, belief propagation [18] and graph cut [19], can produce high-quality disparity maps but require very high computational complexity. Therefore, several semi-global or local stereo methods are generally proposed to achieve efficient implementation [20–23]. However, these semi-global local stereo matching methods still cannot totally solve ambiguity problems, which could come from census transform [21, 22, 24] and local support windows [23]. There are still three main problems need to be solved to improve the precisions for the semi-global stereo matching methods. The determinations of size and shape of local support window should adaptively include more reliable pixels. The sensitivity of intensity in the census transform should be reduced in flat regions that small variations could introduce salt-and-pepper noise in matching cost. Besides, the regular refinement after left-right consistency check cannot unravel the occlusion problems.

To achieve high-precision stereo matching, we propose a semi-global stereo matching system with the trinary cross color (TCC) census transform to reduce sensitivity in smooth region, the two-pass cost aggregation (TPCA) to obtain stable cost, the range winner-take-all (RWTA) to select the robust depth, and the range left-right check (RLRC) to keep the reliable depth. Finally, the triple image-based refinements are also used to further improve the performances. The TPCA combines data term and smooth term together in order to achieve accurate disparity maps in smooth areas and precise object boundaries. The data term is based on the proposed TCC census, which makes raw matching have a better performance than the AD census but with less computation time. A modified RLRC and triple image-based refinements further achieve high-accuracy performance. In this paper, we propose a semi-global stereo matching system based on several techniques, including the TCC census, TPCA, RWTA, and RLRC methods as well as image-based refinements to achieve high-precision depth estimation. The rest of this paper is organized as follows. In Section 2, we first define the stereo matching notations and give a brief overview of the proposed stereo matching system. The details of the framework are described in Section 3. Experimental results to demonstrate the effectiveness of the proposed algorithms are shown in Section 4. Finally, we conclude this paper in Section 5.

2 Local census stereo matching methods

Stereo matching is an active research topic in computer vision. It is one of the important 3D vision methods to recover a dense disparity map from paired images once the inherent ambiguities are properly solved. If the disparity map is generated by simple algorithms, it usually has many discontinuities and occluded areas. If the high-quality disparity map could be produced by complex algorithms [17, 18], the computation time is usually extremely high. Therefore, how to generate the high-quality disparity map with low computing time becomes an important task in 3D vision systems. Figure 1 shows the flow diagram of typical local stereo matching algorithms. In order to reduce the ambiguity problems, local stereo methods commonly aggregate the matching costs of the neighboring pixels in a selected local support window. The local support window should adapt its shape and size to collect the pixels with the same depth. To decrease the sensitivity of intensity, the census transform [21, 22, 25], which relies on the relative ordering of intensities, successfully characterizes the patch structure and achieves good matching property for accurate disparity estimation near depth discontinuities.

With the rectified left and right $W \times H$ color images, with pixels $I_c^l(x,y)$ and $I_c^r(x,y)$, as the inputs of the system. For simplicity, let $p = (x, y)$ indicate the spatial location of the pixel; the left and right images can be simply denoted as $I_c^l(p)$ and $I_c^r(p)$, respectively. For stereo matching, the disparity d should be estimated such that $I_c^l(x,y)$ and $I_c^r(x+d,y)$ become the stereo matched paired pixels, which are also respectively denoted as $I_c^l(p) = I_c^l((x,y))$ and $I_c^r(p,d) = I_c^r((x+d,y))$ for simplicity. For all $W \times H$ pixels, we need to compute all the $W \times H$ disparity values, which are formed as the $W \times H$ disparity map.

To get good raw matching cost, the census transform [22], which converts the intensity as the binarized differences of neighboring pixels, is defined as

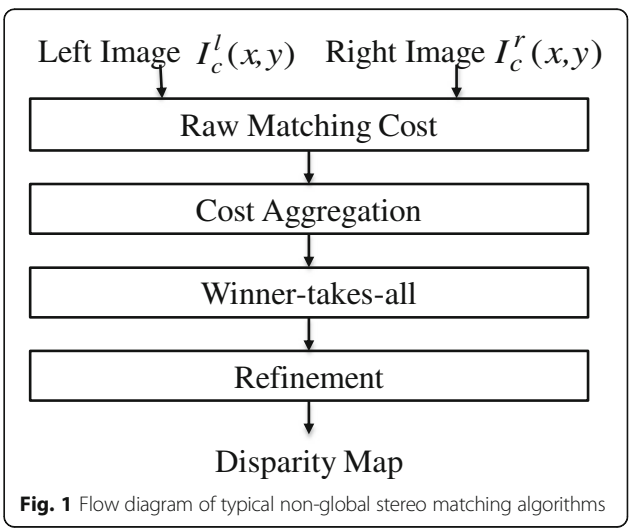

Fig. 1 Flow diagram of typical non-global stereo matching algorithms

$$B_C(p) = \bigotimes_{q \in N(p)} \xi(I(p), I(q)), \tag{1}$$

where $\otimes$ denotes a bitwise catenation operator, and the auxiliary function is defined as

$$\xi(I(p), I(q)) = \begin{cases} 0 & \text{, if } I(q) < I(p) \\ 1 & \text{, otherwise.} \end{cases} \tag{2}$$

where p and q, respectively, denote the positions of the central and surrounding pixels in a selected window $N(p)$, while $I(p)$ and $I(q)$ represent their corresponding intensities of the pixels. The census transform is robust to radiometric distortions and achieves good overall performance in cost representation. However, the census is very sensitive in the flat region that makes the salt-and-pepper noise in matching cost. Besides, the census is obtained from square windows, which could overlay the occlusion areas and expand the boundaries of objects.

3 The proposed stereo matching system

In this paper, for an accurate disparity map generation, we propose a semi-global stereo matching system as shown in Fig. 2. In the proposed system, the matching cost is first computed by the proposed trinary cross color (TCC) census transform, which could attain a reliable measure of pixel dissimilarity. However, the noise generated by raw matching cost still exists though it has been decreased largely by the TCC census transform. To achieve accuracy disparity estimation, we then compute horizontal and vertical smooth terms in pixel-wise fashions to reduce the noise, which could be corrected by the reliable texture edges. Smooth items are mostly used in global stereo matching algorithms, but we combine them with the local stereo matching method such that the proposed algorithm could be treated as a semi-global stereo matching method. After smooth terms computation, we test several census patterns to compute cost aggregation to obtain primary stereo matching. When the different aggregation costs for all possible disparities are obtained, we finally use a range winner-takes-all (WTA) to acquire the disparity map. The disparity map, which still contains many errors, should be further enhanced by the triple image-based refinements stage. In order to identify inconsistent and occluded pixels, we first use the range left-right check (RLRC) to detect them. Once the erroneous pixels are detected, we use color-based voting in the square window to correct them. The voting operation runs iteratively to increase the robustness. Since some artifacts still exist in both left and right disparity maps, we further propose multi-step disparity refinement scheme to achieve the final robust disparity map. With formulation expressions, the details of the key processing units will be described in the following subsections.

3.1 Trinary cross color (TCC) census

With the same bitwise concatenation operation as stated in (1), the trinary function 2-bit format is first proposed as

$$\xi(I(p), I(q)) = \begin{cases} 01, & \text{if } I(q) > I(p) + \rho \\ 10, & \text{if } I(q) < I(p) - \rho \\ 00, & \text{otherwise} \end{cases} \tag{3}$$

to overcome difficulties in finding the correct correspondences in flat areas. In (3), ρ is a selected threshold for reducing the noisy effect and should be proportional to $I(p)$. Figure 3 shows how trinary census works well under noise environments. In the smooth regions, the neighboring pixels show the same intensity should have zero census bits as shown in Fig. 3a. Under noisy environment, the original binary census transform yields very different encoded bits as shown in Fig. 3b, while the trinary census transform produces more consistent encoded bits with only one error as shown in Fig. 3c. Hence, the trinary census transform is more robust to errors than the original one.

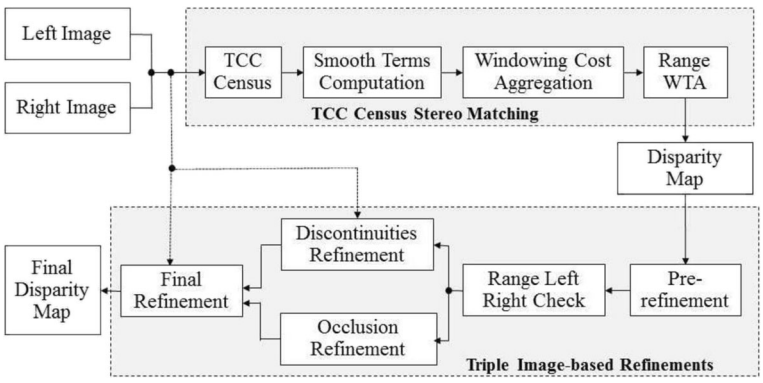

Fig. 2 Framework of the proposed stereo matching system

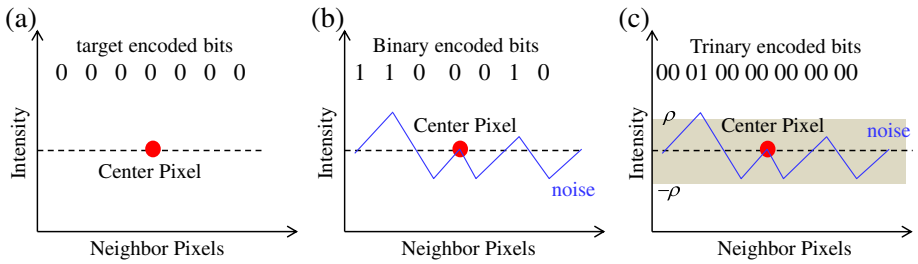

Fig. 3 Census transform results in flat area achieved by **a** target census (without noise), **b** binary census (with noise), and **c** trinary census (with noise)

With trinary census transform, Fig. 4 illustrates four possible patterns. Figure 4a, b has 3×3 and 15×15 square patterns, respectively. Large square windows with more computation generally achieve more reliable results but more likely to be affected by the occlusion area (gray color area) than the smaller ones. To achieve better matching cost and alleviate distorted problem near occultation regions, the rhombus window in Fig. 4c could be used. To further increase the accuracy, the cross-square pattern as shown in Fig. 4d, which covers similar spatial information, is less exposed to the occlusion area as rhombus one. TCC census transform is used to improve the performance of the traditional census matching cost. The entire TCC census transform includes the uses of the trinary census and cross-square pattern. The trinary census could increase the fault-tolerance in flat regions, where the census is very sensitive and makes salt-and-pepper noise in matching cost while the cross-square pattern requires fewer reference points than the large square window and rhombus window, but with higher correctness. In simulations, the TCC census transform, which uses trinary census with the cross-square pattern is suggested to compute the raw matching cost hereafter.

The color information with R, G, and B channels is the most primitive information that we can obtain directly from images; the color similarity $\Delta I_c(p, d)$ between

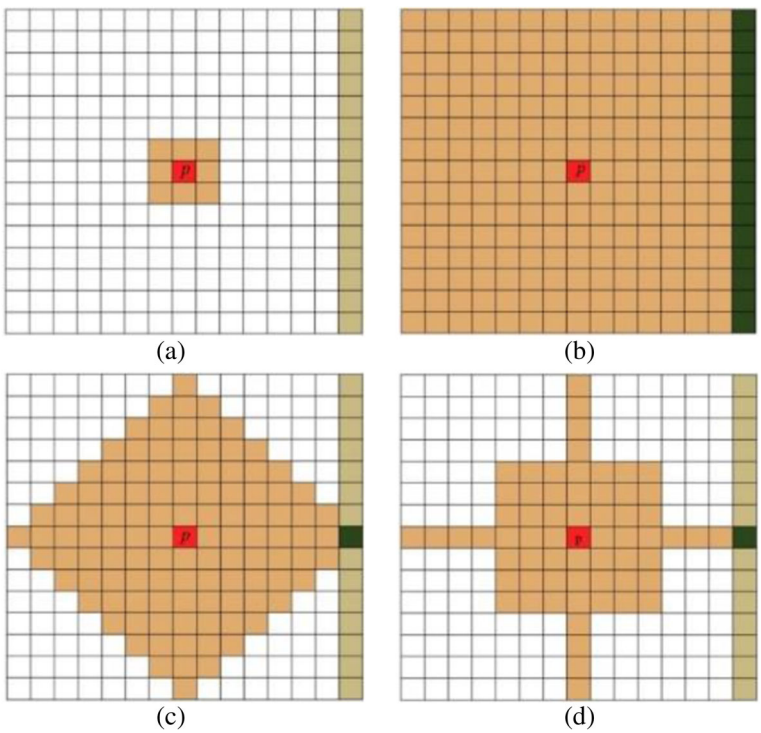

Fig. 4 Four general census patterns. **a** Regular 3×3. **b** Regular 15×15. **c** Rhombus. **d** Cross-square

the pixe1 at p in the left image $I_c^l(p)$, and the pixel at p with disparity d of the right image $I_c^r(p,d)$, can be represented as

$$\Delta I_c(p,d) = \max_{c\in\{R,G,B\}} \left| I_c^l(p) - I_c^r(p,d) \right| \quad (4)$$

where c is the color channel index of the images. The color similarity stated in (4) is insufficient in the raw stereo matching cost for smoothness areas where the census is sensitive. Thus, we use color similarity to detect if we need to use the TCC census cost, which is computed by Hamming distance between TCC census transforms of the pixe1 at p in the left image, $I_c^l(p)$ and the pixel at p with disparity d in the right image, $I_c^r(p,d)$. Thus, the proposed trinary cross color (TCC) census transform cost after normalization is defined as

$$C_{\mathrm{TCC}}(p,d) = \begin{cases} \dfrac{\sum_{c\in\{R,G,B\}} \mathrm{Hamming}\left(B_{\mathrm{TC}}^l(p), B_{\mathrm{TC}}^r(p,d)\right)}{3\times M}, & \Delta I_c(p,d) < T_1 \\ 1, & \text{otherwise} \end{cases} \quad (5)$$

where $B_{TC}^l(p)$ and $B_{TC}^r(p,d)$ are the bit strings of the TCC census transforms of the pixel p in the left and right images, respectively. d is the disparity with respect to the pixel p, T_1 is a threshold to limit the TCC cost, and M is the number of bits in the census window. As shown in Fig. 4d, for example, $M = 16$ (pixels) × 2 (bits/pixel) = 32.

3.2 Smooth processes and cost aggregation

To achieve the semi-global fashion, we first propose to add the smoothness items in row and column directions according to the characteristics of results of initial disparity data items to form a new tectonic energy function model. Then, two levels that improved cross-based cost aggregation based on adaptive support weight are performed to improve the accuracy of disparity map. The smooth terms in the row and column directions could reduce the overall matching error rates, and the modified two-pass cross-based adaptive support weight cost aggregation produces a robust rough disparity maps.

3.2.1 Smooth term computations

The horizontal and vertical direction smooth terms are used to overcome the matching cost errors caused by TCC census raw matching cost.

Figure 5 shows the three-dimensional cost space in horizontal x and vertical y of the image versus disparity d search range. For semi-global disparity estimation, the aim of this disparity space model is first to find the position of minimum disparity cost in horizontal direction from $x = 1$ to $x = W$ by using horizontal iterative smooth term. As shown in Fig. 5, starting at horizontal $x = 1$, we could find the minimum raw matching cost C_{TCC} positions from vertical $y = 1$to $y = H$. The initial horizontal smooth term at $x = 1$ is set as

$$C_{\mathrm{smooth}}^h((1,y),d) = C_{\mathrm{TCC}}((1,y),d);\ y\in[1,\mathrm{H}]. \quad (6)$$

For the horizontal smooth terms at $x \in [2, W]$, we can iteratively compute them as

$$C_{\mathrm{smooth}}^h((x,y),d) = C_{\mathrm{TCC}}((x,y),d) + \lambda \cdot C_{\mathrm{smooth}}^{d_h}((x,y),d) \quad (7)$$

for $x \in [2, \mathrm{W}], y \in [1, \mathrm{H}]$, where the horizontal disparity penalty is given by

$$C_{\mathrm{smooth}}^{d_h}((x,y),d) = \left| d_{\max} - \arg \min_d C_{\mathrm{smooth}}^h((x-1,y),d) \right| \quad (8)$$

and λ is the smooth term parameter; if the value of λ is increased, the occlusion and wrong disparity areas shrink apparently between them, and vice versa.

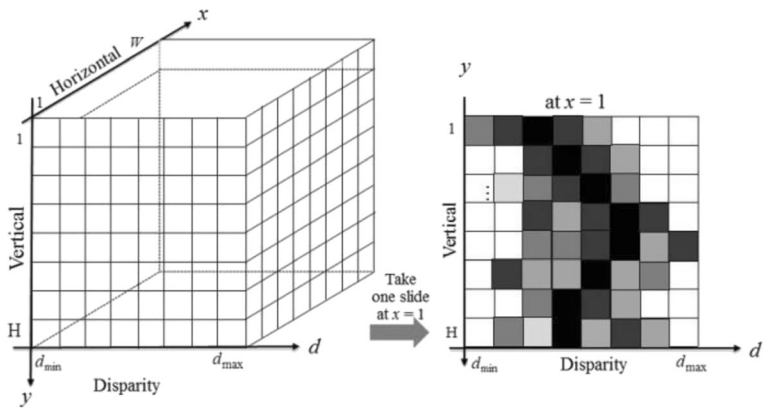

Fig. 5 The diagram of solution using horizontal smooth term

To perform vertical smooth computation, the process will be similar to that of the horizontal smooth computation. However, instead of the TCC census cost, the smooth horizontal cost is used. Initially, for $y = 1$, the vertical smooth term is set as

$$C^{v}_{\text{smooth}}((x,1),d) = C^{h}_{\text{smooth}}((x,1),d)\ ;\ \ x\in[1,\text{W}] \quad (9)$$

for the rest of vertical smooth terms for $y \in [2, \text{H}]$, the iterative computation can be given by

$$C^{v}_{\text{smooth}}((x,y),d) = C^{h}_{\text{smooth}}((x,y),d) + \lambda\cdot C^{d_v}_{\text{smooth}}((x,y),d) \quad (10)$$

for $x \in [1, \text{W}], y \in [2, \text{H}]$, where the vertical disparity penalty term is expressed as,

$$C^{d_v}_{\text{smooth}}((x,y),d) = \left| d_{\max} - \arg\min_{d} C^{v}_{\text{smooth}}((x,y-1),d) \right|. \quad (11)$$

After horizontal and vertical smooth processes, the noises of the disparity map with the TCC census cost can be reduced obviously. Thus, instead of the original TCC census cost $C_{\text{TCC}}(p,d)$, the smooth result $C^{v}_{\text{smooth}}(p,d)$ will be used for stereo matching.

3.2.2 Two-pass cost aggregation

The adaptive cross cost aggregation is used for determination of the rough depth map. The cross window with four arms for pixel p is constructed by considering two measures to find the endpoint pixels of left, right, up, and down arms. The color similarity $\Delta I_c(p)$ in RGB space is defined as

$$\Delta I_c(p) = \max_{c\in\{R,G,B\}} (|I_c(p) - I_c(p_i)|) \quad (12)$$

and the spatial distance $\Delta I_s(p)$ is given by

$$\Delta I_s(p) = |p - p_i| \quad (13)$$

where p is the central pixel for cross-based window generation and I_c is the color intensity of the pixel, where c denotes the R, G, or B color index. In (12) and (13), $i \in [1, L]$, L is the maximum arm length of the cross window. We set the span of left arm r_l as an example. The computation of r_l can be formulated as follows:

$$r_l = \max_{r\in[1,L]} \left(r \prod_{i\in[1,r]} \delta(p,p_i) \right) \quad (14)$$

where $p_i = (x - i, y)$ and $\delta(p, p_i)$ are indicators by gaging color similarity and spatial distance between the pixel p, and p_i as

$$\delta(p,p_i) = \begin{cases} 1, & \Delta I_c(p) \le \tau_k \ \&\ \Delta I_s(p) \le L_k \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

where τ_k and L_k with $k = \{1, 2\}$ are the kth level color similarity threshold and spatial distance threshold, respectively, where $L_1 < L_2$ and $\tau_1 > \tau_2$. After the cross arm construction, the support region for pixel p is developed by merging the horizontal arms of all pixels lying on the vertical arms of p (q for example) as shown in Fig. 6. The proposed two-pass decision cross window allows a more flexible control on the arm length. A larger L_2 contains more pixels for smooth regions but with a stricter τ_2 to guarantee that the arm contains the very similar color regions.

After the construction of cross window, an adaptive support weight function is used for the cost aggregation of pixel p with disparity d as

$$C_{ag}(p,d) = \frac{\sum_{q\in \text{Cross}_p} \omega(p,q) C^{v}_{\text{smooth}}(p,d)}{\sum_{q\in \text{Cross}_p} \omega(p,q)} \quad (16)$$

$$\omega(p,q) = \exp\left(-|p-q|/\gamma_s\right) \quad (17)$$

where Cross_p denotes the detected cross window around p pixel and γ_s is the parameter. If γ_s is increased, C_{ag} will be increased accordingly. In other words, C_{ag} will be weaken if the distance between q and p pixels is larger in the cross window.

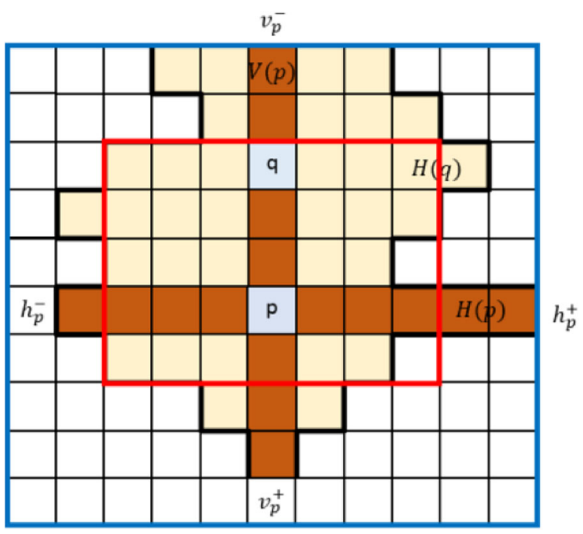

Fig. 6 Two-pass cross window generation in the 1st level region (*red color*) uses a smaller distance threshold, *L*1 and a large color similarity threshold τ1 and uses a larger *L*2 and a smaller τ2 in the 2nd level region (*blue color*)

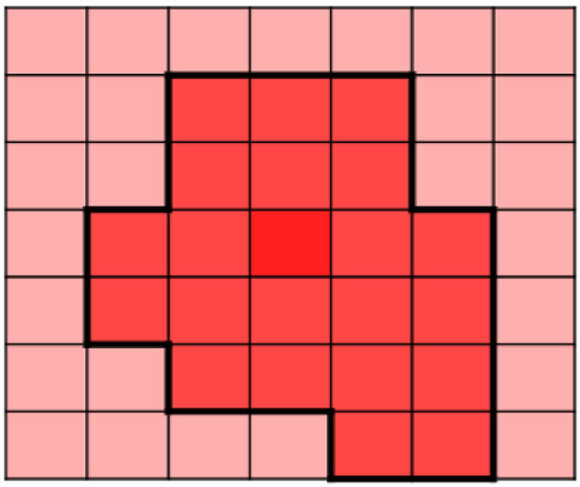

Fig. 7 Example for mismatched areas refinement

3.2.3 Improved WTA for disparity estimation

To estimate disparity map, winner-takes-all (WTA) is then utilized to select the disparity value with the minimum cost evaluated in (17). Since there might exist more than one disparity sharing the same minimum cost for some cases. In smooth regions, the pixels in this case are almost the same that some disparity levels have the same minimum cost. Thus, the traditional WTA cannot achieve a good result in this case. On the contrary, in the repetitive texture areas, there could have same minimum cost in several locations. In order to solve this problem, the WTA procedure should first find all WTA candidates in the increasing ordered as

$$ds(p) = \{d \Big| \arg\min_{d \in D} C_{ag}(p,d)\} = \{d_1, d_2, .d_3 .., d_{N_d}\} \tag{18}$$

where d_{N_d} denotes the candidates of disparity d and $N_d$is the number of disparity levels sharing the same minimum cost. The suggested initial WTA becomes

$$d(p) = \begin{cases} d_1, & \text{if } N_d = 1 \\ d_1, & \text{if } N_d = 2 \text{ and } |d_2 - d_1| = 1 \\ d_2, & \text{if } N_d = 3 \text{ and } |d_3 - d_1| = 2 \\ 255, & \text{otherwise.} \end{cases} \tag{19}$$

If one $\{d\}$, two $\{d, d+1\}$, or three consecutive $\{d-1, d, d+1\}$ disparity levels share with the same minimum cost, the WTA result $d(p) = d$ will be directly adopted in the estimation. However, for $N_d > 3$ or non-consecutive disparities with the same minimum cost, we set the pixel at p as an unstable depth as $d = 255$, which is called as the white hole. In order to fill the white hole, we use cross-based window voting to estimate the disparity as

$$d(p) = d_{\text{vote}}(p) = \arg \max_{d \in D} H_p(d) \tag{20}$$

where $H_p(d)$ is the histogram of the known stable depths in the cross window around p, which was obtained from the first cost aggregation. The depth with the highest histogram bin with the value is selected as the most desirable disparity to fill the white hole.

3.3 Triple image-based disparity refinements

In order to acquire accurate disparity, we have to detect occluded and mismatched areas and refine them first. The pixels in the reference disparity map must have good correspondence to the pixels in the target disparity map. Otherwise, they must be occluded or mismatched.

3.3.1 Occlusion and discontinuities refinement

Let $d_l(x,y)$ and $d_r(x,y)$ be the disparity values in the left and right maps, respectively. The left-right check (LRC) is always used to detect the correct correspondence of the disparities in the left and right depth maps. If the LRC finds $d_l(x,y) = d_r(x - d_l(x,y), y)$, the correct correspondence is detected such that the corresponding disparities will be kept. If the LRC detects $d_l(x,y) \neq d_r(x - d_l(x,y), y)$, we should set the correspondence disparity to be erroneous. To further classify the error pixel as an occluded or mismatched pixel, we further suggest a range LRC as

$$d_l(p) = \begin{cases} d_l(p), & d_l(p) = d_r(p, d_l(p)) \;\&\; |I_l(p) - I_r(p, d_l)| < \delta_1 \\ 255, & d_l(p) = d_r(p, d_l(p) + \sigma) \;\&\; |I_l(p) - I_r(p, d_l(p) + \sigma)| < \delta_2 \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

to detect it for $-d_0 \le \sigma \le d_0$ with $d_0 \ge 1$, where 255 and 0 denote the mismatched and occluded pixels, respectively. In (21), if the left pixel disparity is equal to the pixel disparity with the disparity shift in the right image, these two paired pixels are treated as the correct correspondence. Thus, we keep the original result. If the left

Table 1 The parameters used in the proposed system

Stereo methods	Block	Parameters
TCC census	15 × 15	$\rho = 2$, $T_1 = 20$
TPCA	35 × 35	$\{T_1, T_2, L_1, L_2, \gamma_s\} = \{20, 8, 17, 35, 1\}$,
Refinement	25 × 25	$\{\tau_3, T, L\} = \{15, 20, 35\}$,

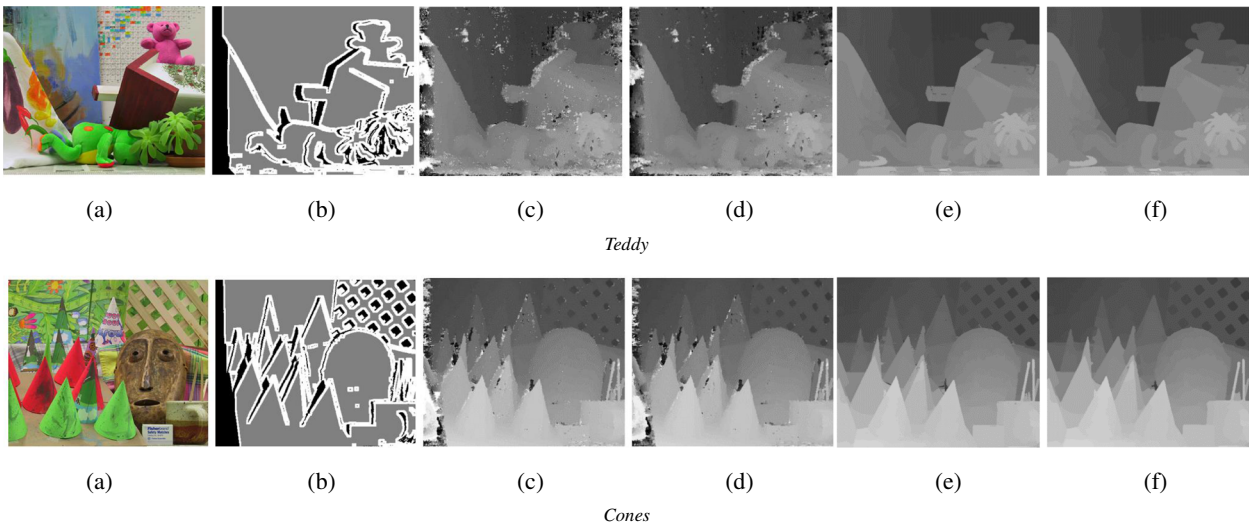

Fig. 8 Results of the proposed algorithms for Teddy and Cones scenes. **a** Color image. **b** Error map. **c** Rough disparity map. **d** Pre-refinement. **e** Multi-step refinement. **f** Final disparity map

pixel disparity finds a matched disparity with a shift of disparity plus a range of $-d_0 \le \sigma \le d_0$ in the right image, the erroneous pixel will be marked as mismatched pixel with 255. If the pixel cannot find the matched pixel either with the disparity shift or with a range of disparity shift, we set this pixel as an occluded pixel with 0.

When the occluded and mismatched areas are detected, we use different methods based on the corresponding color image to refine them. For the occluded pixel, we adopt the lowest stable disparity around it for the refinement since it most likely comes from the background. The refinement is

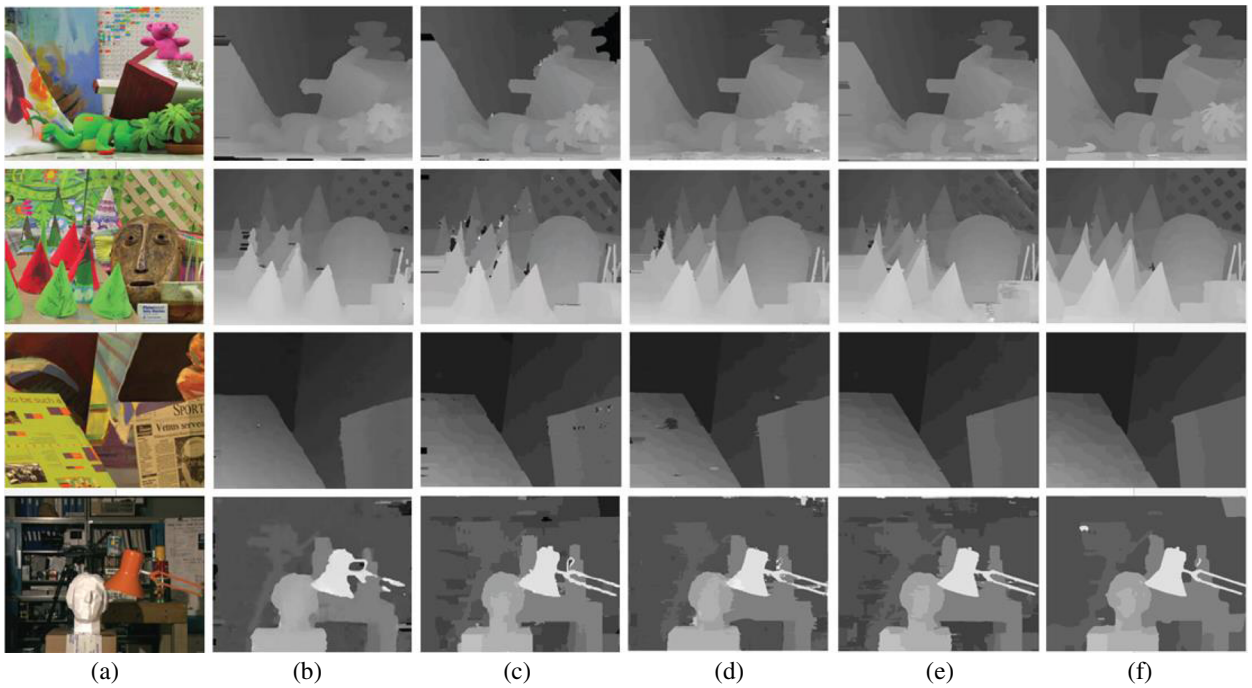

Fig. 9 Results of the estimated disparity maps of **a** color images achieved by **b** census-based semi-global, **c** cross-based local, **d** combined cost-based local, **e** belief propagation-based global, and **f** proposed stereo matching methods. *Top* to *bottom*: *Teddy*, *Cones*, *Venus*, *Tsukuba*

Table 2 Characteristics of Middlebury 2014 stereo datasets

Characteristic	Datasets
Normal	Adirondark, Motorcycle, Piano, Pipes, Playroom, Playtable, Recycle, Shelves, and Teddy
Light	ArtL, MotorcycleE, and PianoL
Large disparity	Jadeplant and Vintage
Angle moving	PlaytableP

$$d_l(p) = \min\{d_l(x-1,y), d_l(x,y-1), d_l(x+1,y), d_l(x,y+1)\},$$
$$\text{if } I_l(x,y) \in \{I_l(x-1,y), I_l(x,y-1), I_l(x+1,y), I_l(x,y+1)\} \quad (22)$$

where $d_l(x,y)$ is the occluded pixel if its four surrounding pixels have reliable disparities. With iterative refinements, the occluded pixels (black holes) will be successfully refined with the background. For the mismatched pixels (white holes), we use the window voting based on the corresponding color image for the largest proportion stable pixels selection as

$$d(p) = d_{\text{vote}}(p) = \arg\max H_W(d) \quad (23)$$

where $H_W(d)$ is the histogram of the stable and color-matched depths around p in the $K \times K$ voting window W, where the color-matched pixel is defined as

$$C(p_i) = \begin{cases} 1, & \Delta I_c(p_i) \le \tau_3 \\ 0, & \text{otherwise} \end{cases} \quad (24)$$

with the color similarity as

$$\Delta I_c(p_i) = \max_{c \in \{R,G,B\}} (|I_c(p) - I_c(p_i)|) \quad (25)$$

and τ_3 is a color similarity threshold, for example, the circled area in Fig. 7 shows the similar color space for the correct pixel voting.

3.3.2 Final disparity map refinement

There are still some noises and wrong disparities in the disparity map. We use the cross-based window voting for the disparity with the maximum number in this area to refine them. The cross window is constructed with the same method in section B, and the disparity of stable pixels with maximum number in this area is selected to replace it. Finally, a 3×3 median filter is used to obtain the smoothness disparity map.

4 Experimental results

The experimental evaluation of the proposed stereo matching system is performed by using Middlebury datasets [26]. In Section 4.1, we first show the disparity maps achieved by the proposed methods stage by stage to analyze the improvement in each step. In Section 4.2, we then compare the proposed stereo matching system to the other well-known methods. The disparity maps, which are generated by the proposed and compared methods, will be exhibited.

4.1 Performance evaluation of the proposed algorithm

We use the 2001 and 2003 datasets suggested in the Middlebury for evaluation of the main algorithms in the proposed stereo matching system. For Middlebury datasets, the dimensions of images Tsukuba, Venus, Teddy, and Cones are 384×288, 434×383, 453×375, and

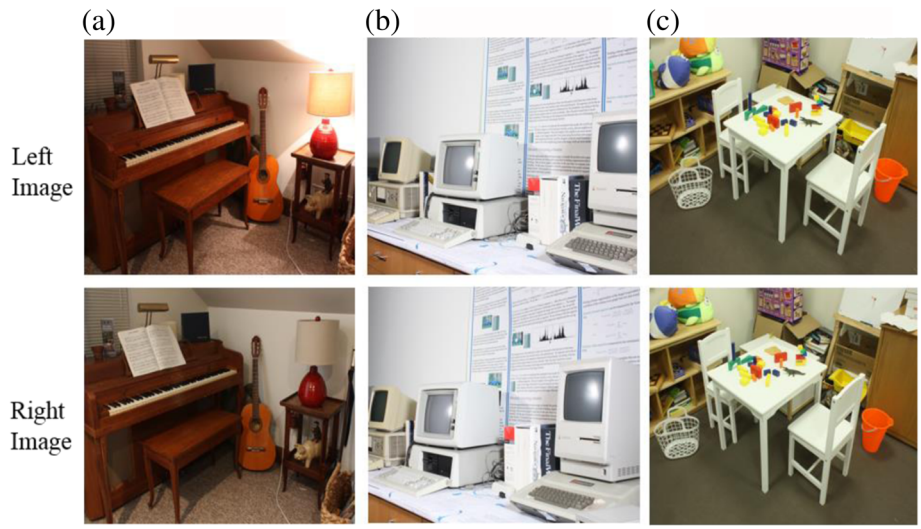

Fig. 10 Different characteristics of **a** light, **b** large disparity, and **c** angle moving conditions in Middlebury 2015 stereo datasets

450 × 375 with the disparity level of 15, 19, 59, and 59, respectively. We applied the proposed algorithms step by step to calculate the disparity results. We chose the experiment's parameters empirically and kept them constant as shown in Table 1. After simulations, Fig. 7 shows the color images, ground truth of error maps, rough disparity maps, pre-refinement, multi-step refinement, and final disparity maps.

With TCC census matching cost computation, which considers the trinary of the differences of three color components in a cross window, we use the TPCA to achieve rough disparity images as shown in Fig. 8c for both Teddy and Cones scenes. We learnt that the TPCA greatly helps to achieve considerable good results. Then, the RWTA algorithm further improves the initial rough disparity maps by detecting the white holes. To fill the white holes, the cross-based window voting method is used by referring to the color image, and the results after pre-refinement are shown in Fig. 8d. However, there are still some ambiguous regions in the disparity maps to make them not accurate enough due to occlusion regions that occurred and discontinuities that mismatched. The occlusion areas could be corrected by the background disparities, while discontinuities that mismatched should be corrected by the most similar pixels around it. Therefore, we further use cross-based window voting for the disparity with the maximum occurrence to refine them as shown in Fig. 8e. Finally, the median filter is used to obtain the smoothness disparity maps as exhibited in Fig. 8f. The proposed stereo matching system achieves considerable good performance in depth estimation.

4.2 Performance comparisons

In this subsection, we compare the proposed system to four related methods, which are census-based semi-global stereo matching [21], cross-based local stereo matching [15], combined cost-based local stereo matching [27], and belief propagation-based global stereo matching [18]. In Fig. 9, the results show that the proposed method yields competitive results comparing to these four methods. By observing Fig. 9b, f, the disparity maps produced by the census-based and the proposed semi-global stereo matching methods are very close. However, in Fig. 9b, the occlusion regions still show unsolved results near some object boundaries. Moreover, there also have some chaotic results near strong edges in complex texture areas. It is evident that the proposed system achieves more accurate performance in these areas. From Fig. 9c, f, we find that the results obtained

Fig. 11 Results in Middlebury 2015 stereo datasets. **a** Color image. **b** Ground truth. **c** Estimated disparity map

Table 3 Rank and analyzed performances of the proposed system with normal, light, large disparity, and angle moving conditions

Performances	RMS disparity error		Average absolute error (Avgerr.)		99% error quantile (A99)	
Datasets	RMS	Rank	Avgerr.	Rank	Avgerr.	Rank
Normal						
Adirondark	16.7	13	6.4	12	97.6	15
Motorcycle	19.8	11	5.7	15	121.0	9
Piano	9.5	4	5.1	11	39.9	3
Pipes	32.7	15	12.6	13	153.0	15
Playroom	14.7	3	7.0	8	65.5	3
Playtable	19.0	3	10.1	5	63.7	2
Recycle	10.3	9	4.9	15	40.0	4
Shelves	21.1	12	10.6	11	91.4	15
Teddy	7.81	7	4.1	15	35.0	4
Average	16.8	8	7.4	11	78.5	7
Light						
ArtL	47.8	20	31.1	23	154.0	17
MotorcycleE	78.8	23	55.7	23	206.0	19
PianoL	61.6	22	35.4	21	193.0	23
Large disparity						
Jadeplant	112.0	19	51.4	19	444.0	21
Vintage	30.8	10	14.0	9	96.9	6
Angle moving						
PlaytableP	12.2	12	7.0	16	48.1	6
Average	57.2	17	32.5	18	190.3	15

by the proposed system are better than those achieved by the cross-based local stereo matching method. With the proposed system, most disparity values in occlusion regions near the object edges are correctly retrieved. From Fig. 9d, f, the disparity maps generated by the cross-based local stereo matching method have many inaccurate areas in foreground objects; inaccurate areas are often produced by the cost aggregation step of the local-based method in the smooth areas of the color image. On the other hand, from Fig. 9e, f, the disparity maps generated by the belief propagation-based global stereo matching method are more correct in the object. However, computational time and computational complexity of global-based methods are too higher compared with those of the other methods, which are not conducive to hardware implementation and real-time computing system.

Table 2 shows the characteristic of Middlebury 2014 stereo datasets. Middlebury has used these datasets for the ranks of different stereo matching methods in 2015. There are 15 pairs of stereo matching datasets: nine normal stereo datasets, three stereo datasets with different light conditions, two large disparity (more than 150) stereo datasets, and one angle moving stereo dataset, as shown in Fig. 10.

Figure 11 shows the disparity maps estimated by the proposed stereo matching system, and Table 3 exhibits the rank and analyzed performances of root mean square (RMS) disparity error, the rank and analysis of average absolute error (Avgerr.) and disparity error, and the rank and analysis of 99% error quantile (A99) of the proposed stereo matching system. The proposed system achieves more accurate results comparing to other local stereo matching methods, especially in the normal stereo datasets. However, this is not ideal in light, large disparity, and angle moving stereo datasets, while other methods are as the same processing properties in these datasets.

In summary, the proposed stereo matching method achieves better disparity estimation quality than the methods proposed in [15, 18, 21] and [27] in Middlebury datasets, especially in the normal stereo datasets. For non-ideal conditions in light, large disparity, and angle moving stereo datasets, the proposed and the other

Table 4 Execution times of the proposed and other stereo matching methods

Methods	[21]	[15]	[27]	[18]	Proposed method
Execution time(s)	8546.28	2.47	351.65	1627.56	298.18

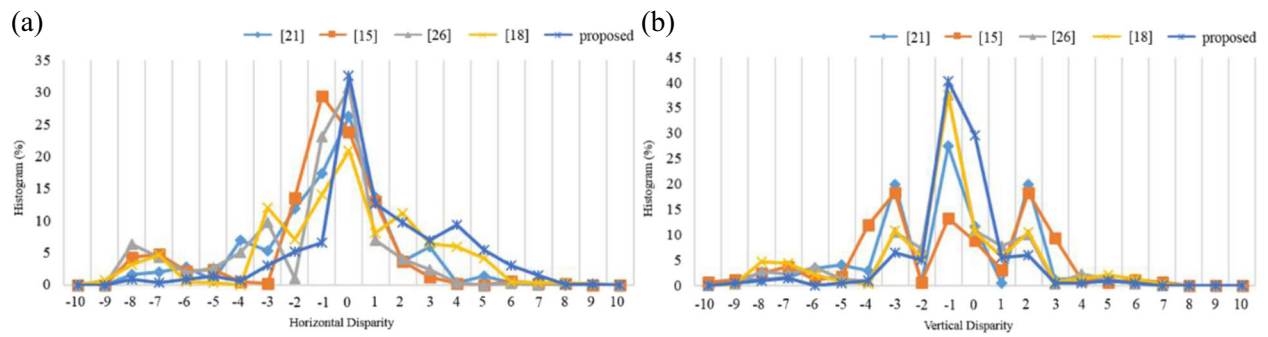

Fig. 12 Results of the estimated disparity map quality metric. **a** Horizontal disparit. **b** Vertical disparity

methods have the same tendency in these datasets. Table 4 shows the total execution times required by the proposed and other stereo matching methods. In general, the filter window sizes used by the local-based stereo matching methods are usually smaller than those used in the global and semi-global-based stereo matching methods. Thus, the execution times of local-based methods are less than those of the global- and semi-global-based methods.

In cross-based local stereo matching [15], the simulation results shown in Fig. 9c, f exhibit that the proposed method is better than the cross-based local stereo matching method. The proposed method correctly retrieves the most disparity values in occlusion regions near the object edges; on the other hand, the proposed method acquires a less execution time and achieves higher performance than the existed semi-global stereo matching methods. The proposed stereo matching system is about 28 times faster than the census-based semi-global methods, where the experiments are carried on an Intel Core i7-4770 CPU computer with a 12-GB RAM and tested on the Matlab platform (Version R2013a). It is noted that the computation time can be further improved with graphics processing units (GPU) parallel computation. The histograms of horizontal disparity and vertical disparity are shown in Fig. 12. In order to obtain an objective evaluation of the estimated disparity map quality, the estimated disparity map quality metric suggested in [28] is used for comparisons. Figure 12a shows that the estimated disparities of the proposed method have smaller disparities in the left half of the histogram, representing less visual fatigue. Figure 12b exhibits that the estimated disparities of the proposed method focus on the central region of the histogram which means less visual fatigue. The proposed method with the trinary cross color (TCC) census transform reduces the sensitivity in smooth regions and that with the two-pass cost aggregation (TPCA) obtains the stable cost. Thus, in the proposed method, the disparity values of the smooth regions can be correctly calculated and successfully limited to a narrow range without causing the noise and large occlusion regions.

5 Conclusions

In this paper, a semi-global stereo matching system based on improved TCC census cost, TPCA, and triple image-based refinements is proposed. The TPCA combines data term and smooth term together in order to achieve accurate disparity maps in smooth areas and precise object boundaries. The data term is based on the proposed trinary cross color (TCC) census, which makes raw matching have a better performance than the AD census but with less computation time. The TPCA method with the smooth term iteratively removes the noise caused by TCC census raw matching. The cross-based cost aggregation with two-pass and adaptive support weights is performed to make accurate results in the same color areas. A modified range left-right check (RLRC) and multi-step refinements further achieve high-accuracy performance. The detection of the occluded and mismatched pixels helps us to apply the corresponding method to refine them. Several extended experimental results based on multiple stereo pairs prove the efficiency of the proposed approach compared to the related corresponding method with respect to disparity estimation problems. Two steps of disparity estimation and disparity map refinement increase computational cost mainly caused by cost aggregation in multiple loops. However, the proposed TCC census, TPCA, and triple image-based refinements help to achieve more accurate disparity map estimation in comparison with other related methods. For real-time applications, the GPU or VLSI implementation of the system should be further studied. In addition, the improvement of the subpixel level accuracy of depth estimation could be also investigated to attain better virtual view syntheses and possibly be used for the free-view 3D video generation.

Funding

This work was supported in part by the National Science Council of Taiwan, under Grant MOST 105-2221-E-006 -065 -MY3.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
TAC carried out the image processing studies, participated in the proposed system design, and drafted the manuscript. XL carried out the figure design and adjustment parameters. JFY conceived of the study and participated in its design and coordination and helped in drafting the manuscript. All authors read and approved the final manuscript.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 14 September 2016 Accepted: 17 March 2017
Published online: 31 March 2017

References
1. C Faria, W Erlhagen, M Rito, E Demomi, G Ferrigno, E Bicho, Review of robotic technology for stereotactic neurosurgery. IEEE Trans on Biomedical Engineering **8**, 125–137 (2015)
2. G Gioioso, G Salvietti, M Malvezzi, D Prattichizzo, Mapping synergies from human to robotic hands with dissimilar kinematics: an approach in the object domain. IEEE Trans on Robotics **29**(4), 825–837 (2013)
3. ZF Zhang, YC Xu, JF Liu, Design of intelligent vehicle control system of self-directed. IEEE Control Decis Conf 2748–2751 (2012)
4. J Cai, Integration of optical flow and dynamic programming for stereo matching. IET Image Process **6**(3), 205–212 (2012)
5. HM Wang, YH Chen, JF Yang, A novel matching frame selection method for stereoscopic video generation. IEEE Multimedia and Expo Conf 1174–1177 (2009)
6. EG Mora, J Jung, M Cagnazzo, B Pesquetpopescu, Initialization, limitation, and predictive coding of the depth and texture quadtree in 3D-HEVC. IEEE Trans Circuits Systems Video Technol **24**(9), 1554–1565 (2014)
7. G Tech, K Wegner, Y Chen, S Yea, *3D HEVC Test Model 3* (Document: JCT3VC1005, Geneva, 2013)
8. QH Nguyen, MN Do, SJ Patel, Depth image-based rendering with low resolution depth. IEEE Image Process Conf 553–556 (2009)
9. CH Hsia, Improved depth image-based rendering using an adaptive compensation method on an autostereoscopic 3-D display for a Kinect sensor. IEEE Trans on Sensors **15**(2), 994–1002 (2015)
10. YA Sheikh, M Shah, Trajectory association across multiple airborne cameras. IEEE Trans Pattern Anal Mach Intell **30**(2), 361–367 (2008)
11. H Hirschmuller, Stereo vision is structured environments by consistent semi-global matching. IEEE Computer Vision and Pattern Recognition Conf **2**, 2386–2393 (2006)
12. SB Kang, R Szeliski, J Chai, Handling occlusions in dense multi-view stereo. IEEE Computer Vision Patter Recognition Conf **1**, 103–110 (2001)
13. VQ Dinh, CC Pham, JW Jeon, Matching cost function using robust soft rank transformations. IET Image Process **10**(7), 561–569 (2016)
14. JB Lu, S Rogmans, G Lafruit, F Catthoor, Stream-centric stereo matching and view synthesis: a high-speed approach on GPUs. IEEE Trans Circuits Systems Video Technol **19**(11), 1598–1611 (2009)
15. K Zhang, JB Lu, G Lafruit, Cross-based local stereo matching using orthogonal integral images. IEEE Trans Circuits Syst Video Technol **19**(7), 1073–1079 (2009)
16. A Hosni, M Bleyer, C Rhemann, M Gelautz, C Rother, Real-time local stereo matching using guided image filtering. IEEE Multimedia and Expo Conf 1–6 (2011)
17. D Schartein, R Szeliski, A taxonomy and evaluation of dense two-frame stereo *correspondence* algorithms. Int J Comput Vis **47**(1–3), 7–42 (2002)
18. J Sun, N Zheng, H Shum, Stereo matching using belief propagation. IEEE Trans Pattern Anal Mach Intell **25**(7), 787–800 (2003)
19. Y Boykov, O Veksler, R Zabih, Fast approximate energy minimization via graph cuts. IEEE Trans Pattern Anal Mach Intell **23**(11), 1222–1239 (2001)
20. H Hirschmuller, D Scharstein, Evaluation of stereo matching costs on images with radiometric differences. IEEE Trans Pattern Anal Mach Intell **31**(9), 1582–1599 (2009)
21. M Humenberger, T Engelke, W Kubinger, A census-based stereo vison algorithm using modified semi-global matching and plane fitting to improve matching quality. IEEE Comput Vis Patter Recognition Conf 77–84 (2010)
22. KR Bae, HS Son, J Hyun, B Moon, A census-based stereo matching algorithm with multiple sparse windows. IEEE Ubiquitous Future Networks (ICUFN) Conf 240–245 (2015)
23. A Fusiello, V Roberto, E Trucco, Efficient stereo with multiple windowing. IEEE Comput Vis Patter Recognition Workshop Conf 858–863 (1997)
24. R Zabih, J Woodfill, Non-parametric local transforms for computing visual correspondence. European Comput Vis Conf **2**, 151–158 (1994)
25. X Mei, X Sun, MC Zhou, SH Jiao, HT Wang, XP Zhang, On building an accurate stereo matching system on graphics hardware. IEEE Comput Vis Workshops Conf 467–474 (2011)
26. http://vision.middlebury.edu/stereo. Accessed 29 Mar 2017, Middlebury stereo vision page [Online]
27. J Jiao, R Wang, W Wang, S Dong, Z Wang, W Gao, Local stereo matching with improved matching cost and disparity refinement. IEEE Trans Multimedia **21**(4), 16–27 (2014)
28. D Kim, D Min, J Oh, S Jeon, K Sohn, Depth map quality metric for three-dimensional video. IS&T/SPIE Electronic Imaging Conf 723719 (2009)