Int J Comput Vis (2016) 120:256–271
DOI 10.1007/s11263-016-0910-9

Occlusion-Aware Stereo Matching

Jintao Xu[1] · **Qingxiong Yang**[2] · **Zuren Feng**[1]

Received: 15 August 2015 / Accepted: 15 April 2016 / Published online: 29 April 2016

Abstract Stereo vision systems with additional flash/no-flash cues have been demonstrated to be robust to depth discontinuities. The ratio of a flash and no-flash image pair naturally provides additional scene depth information and thus can serve as a strong cue for preserving depth discontinuities. However, existing solution simply uses ratio as the guidance to perform matching cost aggregation and thus is still vulnerable to occlusions. Inevitable misalignment of flash and no-flash images due to camera and/or scene motion remains unsolved as well. This paper investigates into these two problems. An occlusion detection approach is derived based on foreground/background extraction. Matching cost computed in the occluded regions (which is useless and harmful) is thus discarded so that reliable information from non-occluded regions can be easily propagated in. The foreground, occlusion and depth estimation is modeled in a uniform framework base on Expectation-Maximum. The proposed solution is evaluated using both indoor and outdoor data sets, showing clear improvement over the state-of-the-art methods.

Keywords Stereo matching · Cost aggregation · Foreground segmentation · Flash photography

Communicated by Long Quan.

✉ Qingxiong Yang
liiton.research@gmail.com

[1] Institute of Systems Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[2] School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui, China

1 Introduction

Accurate depth acquisition is one of the fundamental research areas in computer vision. The depth information are very useful to many vision tasks like scene understanding (Kaehler and Reid 2013; Xiong et al. 2009), 3D object modeling (Rothganger et al. 2006; Ye et al. 2012), robot vision (Murray and Little 2000) and tracking (Prisacariu and Reid 2012; Ren et al. 2013). Depth images can be computed using active depth sensing techniques or passive stereo. For instance, laser scanner (Riegl vz 1000 scanner, http://www.riegl.com/nc/products/terrestrial-scanning/produktdetail/product/scanner/27/, time-of-flight (ToF) (Softkinetic depth sensor, http://www.softkinetic.com/Products/DepthSenseCameras, 2015) and structured-light (Zhang 2012) depth cameras. There are other depth sensing systems based on different types of techniques including light field (Chen et al. 2014; Yu et al. 2013), XSlit cameras (Ye et al. 2013) and mixed camera types (Bastanlar et al. 2012). However, they are less popular. Some of the commercial depth cameras can even capture depth images in real time with sufficient depth quality. However, most popular depth sensing techniques have limitations as summarized in Table 1 and some failure cases are presented in Fig. 1.

A state-of-the-art laser scanning system can be extremely accurate and of high resolution. However, it is very expensive and extremely slow. For instance, the Riegl vz 1000 scanner, http://www.riegl.com/nc/products/terrestrial-scanning/produktdetail/product/scanner/27/ takes about half an hour to capture a 4343×2848 depth image. Additionally, the foreground depth will be blended with the background because the width of the adopted laser beam cannot be infinitely small in practice. Such a failure case is presented in the closeups in Fig. 1a. Figure 1b, c show that Structured-Light and ToF depth sensing techniques are not suitable for thin-structured objects due to limited sensor resolution. Passive stereo cannot

Table 1 Popular depth sensing techniques

	Advantages	Disadvantages
Laser scanner	Accurate; High resolution	Slow and expensive; Blurred depth edges.
Structured-light	Accurate	Low resolution; Limited in depth range; Sensitive to occlusion
Time-of-flight	Accurate; Robust to occlusion	Low resolution; Limited in depth range
Traditional passive stereo (Hosni et al. 2013)	High resolution.	Sensitive to occlusion; Cannot distinguish color Edges and depth edges
Flash stereo (Zhou et al. 2012)	High resolution; Robust to depth edges	Sensitive to occlusion; Sensitive to motion; Require camera flash
Proposed	High resolution; Robust to occlusion, Depth edges and motion.	Require camera flash.

distinguish color edges and depth edges and thus is sensitive to occlusions and depth discontinuities as demonstrated in Fig. 1d. Flash stereo (Zhou et al. 2012) uses an additional flash stereo pair. It can provides approximated depth information and thus can better preserve depth discontinuities. However, is still vulnerable to occlusions and inevitable misalignment of flash and no-flash images due to camera motion and scene motion. The depth sensing technique proposed in this paper aims at solving these limitations.

A typical stereo matching algorithm consists of a matching cost computation step followed by a cost aggregation and/or a disparity optimization step (Scharstein and Szeliski 2002). Cost aggregation methods usually perform a winner-takes-all (WTA) operation for each pixel after aggregating matching costs from other pixels, while optimization methods strive to infer a global optimal disparity value for each pixel.

Thanks to the Middlebury stereo benchmark (Scharstein and Szeliski, http://vision.middlebury.edu/stereo/eval/), significant progress on stereo matching has been achieved in the past several decades. However, there are still some crucial issues which are intractable in practice. For instance, distinguishing depth and color edges for preserving depth discontinuities and inferring depth inside large occlusions. Most of the state-of-the-art solutions use edge-aware aggregation/filtering techniques (Yoon and Kweon 2006; He et al. 2013; Hosni et al. 2013; Yang 2012; Ma et al. 2013) based on the assumption that depth discontinuities co-exist with color edges. Nevertheless, this assumption is not suitable for textured scenes.

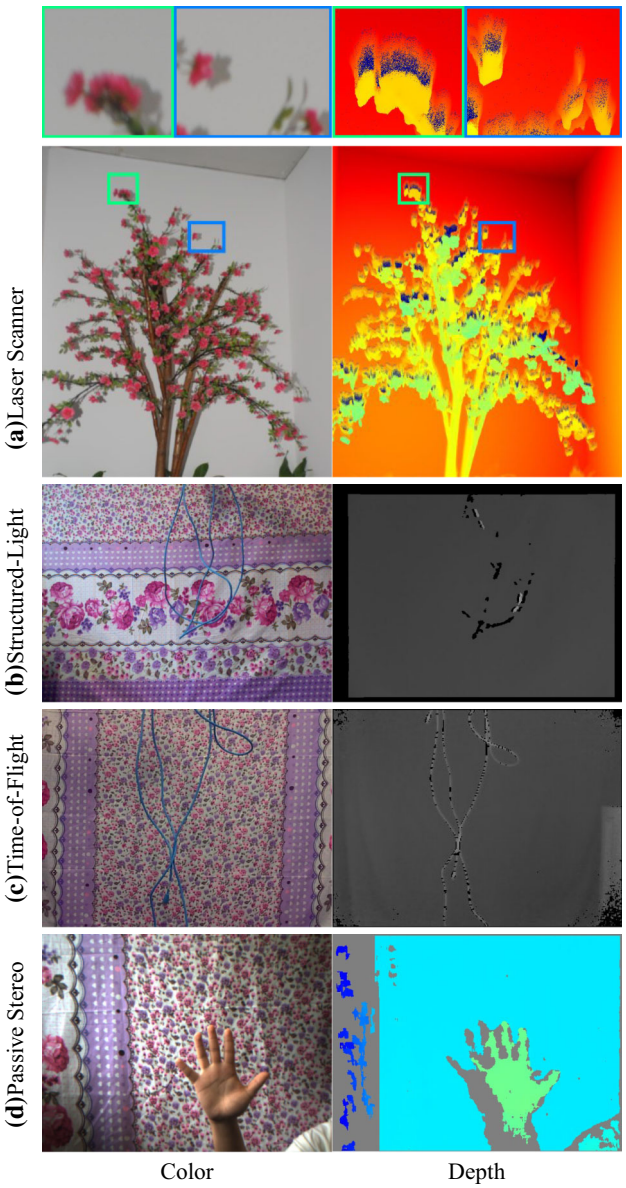
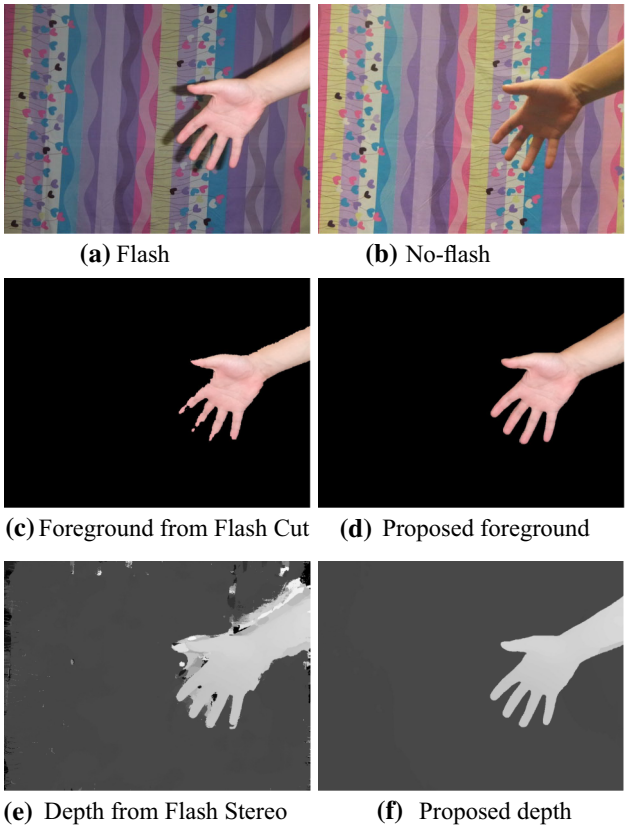


Fig. 1 Limitation of existing depth sensors/cameras. From **a** to **d**: Color images and depth images captured by a state-of-the-art high-resolution (but extremely slow) laser scanner (Riegl vz 1000 scanner, http://www.riegl.com/nc/products/terrestrial-scanning/produktdetail/product/scanner/27/), Kinect (Structured-Light) (Zhang 2012), Soft-Kinetic (Time-of-Flight) (Softkinetic depth sensor, http://www.softkinetic.com/Products/DepthSenseCameras, 2015) and Point Grey Bumblebee2 stereo camera (Point-gray stereo camera, http://www.ptgrey.com//bumblebee2-firewire-stereo-vision-camera-systems, 2015), respectively. As can be seen in the closeups in (**a**), the foreground depth from a laser scanner will be blended with the background because the width of the adopted laser beam cannot be ignored. **b**, **c** show that Structured-Light and Time-of-Flight depth sensing techniques are not suitable for thin-structured objects due to limited sensor resolution, and apparently passive stereo is non-robust around depth edges as demonstrated in (**d**). The proposed depth sensing technique aims at solving these limitations

The optimal guidance for matching cost aggregation is indeed the depth/disparity image to be computed. A more robust way to guide the weighting scheme is thus to utilize additional cues which can provide reliable depth discontinuity information. It is a great success in digital image matting when (Sun et al. 2006) propose to add a flash image cues for matting. It is based on the observation that most noticeable difference between flash and no-flash image is the foreground object if the background is relatively far away. The use of the additional flash image simplifies the image matting problem. An accurate foreground/background layer extraction method named **FlashCut** is later proposed by Sun et al. (2007) based on the same observation. It is more practical than the matting solution as the flash and no-flash image misalignment problem is considered. Zhou et al. (2012) further prove that the ratio map of a flash and no-flash image pair varies based on the surface normal and object distance. This method is referred to as **Flash Stereo** in this paper. The ratio map is thus used as the guidance (of a joint bilateral filter) for matching cost aggregation and has been demonstrated to impressively outperform the traditional color guidance. However, Zhou et al. (2012) merely uses the ratio map as the guidance to perform matching cost aggregation and thus is still vulnerable to occlusions. Inevitable misalignments of flash and no-flash images due to camera motion and scene motion are not considered either.

This paper proposes a new framework for the integrated estimation of the foreground, occlusion and depth using flash/noflash stereo image pairs. Considering these problems separately has clear limitations and fails to take advantage of their complementary nature. For instance, **FlashCut** background modeling technique (Sun et al. 2007) relies on the difference between the flash and noflash images and thus the performance is directly related to the surface albedo and the distance of the background object as shown in Fig. 2a, c. These limitations can be successfully suppressed with depth estimates from stereo matching as demonstrated in Fig. 2d. Matching cost aggregation based on only the ratio of the flash/noflash images is vulnerable to camera/scene motion. The use of ratio does not sufficiently utilize the rich information containing in a flash/noflash image pair. For instance, the matching costs computed inside occlusions are mostly outliers and should be excluded from cost aggregation. Otherwise the depth estimates of the occlusions are likely to be incorrect as demonstrated in Fig. 2e. This paper derives an occlusion map from background modeling (using the flash/noflash pair) and integrate it with the foreground/background segmentation result for edge-aware and occlusion-aware stereo matching as shown in Fig. 2f. The combination of background modeling and cost aggregation also solves the inevitable misalignments of flash and no-flash images problem in **Flash Stereo** (Zhou et al. 2012).



(a) Flash **(b)** No-flash

(c) Foreground from Flash Cut **(d)** Proposed foreground

(e) Depth from Flash Stereo **(f)** Proposed depth

Fig. 2 Some limitations of FlashCut (Sun et al. 2007) and Flash Stereo (Zhou et al. 2012) are visualized in **c** and **e**, respectively. The performance of FlashCut drops when the background gets close to the foreground (as the background will be also changed significantly by flash) while Flash Stereo is sensitive to occlusions and motions. The proposed system is more robust to these limitations as demonstrated in **d** and **f**

By modeling the foreground, occlusion and depth estimation in a uniform framework base on Expectation-Maximum, a robust depth sensing system is proposed in this paper. It has been evaluated using sufficient indoor and outdoor data sets, and both visual and numerical comparison demonstrate clear improvement over the state-of-the-art.

2 Proposed Solution

This section proposes a uniform framework to jointly estimate foreground, occlusion and depth with a flash and no-flash stereo image pair. A brief review of the FlashCut background modeling technique (Sun et al. 2007) and Flash stereo (Zhou et al. 2012) is presented in Sect. 2.1. A detailed Expectation-Maximization (EM) based parameter estimation model is then introduced to estimate the foreground, occlusion and depth from Sects. 2.2 to 2.5. The whole system pipeline is briefly summarized in Algorithm 1.

Algorithm 1 Occlusion-aware stereo matching.

Initialize
- foreground probabilities α_l and α_r with flash differences between flash and no-flash images;
- disparity maps with Hosni et al. (2013);
- foreground/background layers with Sun et al. (2007).

repeat
- Estimate parameters α_l and α_r with EM optimization (Sect. 2.3);
- Improve foreground/background layers (Sect. 2.4);
- Improve disparity map (Sect. 2.5).

until Convergence.

2.1 Motivation

Flash cue is used in FlashCut (Sun et al. 2007) for foreground segmentation. FlashCut (Sun et al. 2007) assumes that the foreground will be significantly brightened by the flash while the background appearance change is very small. This assumption is valid when the background is sufficiently far away from the foreground. Meanwhile, FlashCut (Sun et al. 2007) handles misalignments caused by camera shake and/or scene motion. However, real scenes may contain white/black objects which are hard to be brightened by flash. Slanted surfaces (which are not perpendicular to flash source) can not be significantly brightened by flash neither. Indoor scenes are also quite difficult because background will be also affected by flash as shown in Fig. 2a, b. These limitations may reduce the performance of FlashCut in practice as demonstrated in Fig. 2c. The foreground fingers cannot be fully detected by FlashCut as the background is close to the fingers.

On the other hand, Flash Stereo (Zhou et al. 2012) prove that the ratio of a flash and no-flash image pair varies based on the surface normal and object distance and is thus used as the guidance for matching cost aggregation. It has been demonstrated to impressively outperform the traditional color guidance. However, it is still vulnerable to occlusions. Inevitable misalignments of flash and noflash images due to camera motion and scene motion are not considered either. These limitations lead to the incorrect depth estimates in Fig. 2e.

This section analyzes the complementary nature of the FlashCut (Sun et al. 2007) and Flash Stereo (Zhou et al. 2012) and proposes to utilize the flash cue and depth cue simultaneously. Accurate foreground/background layers can provide a good guidance to tackle occlusions in stereo matching using the proposed occlusion detection technique (as detailed in Sect. 2.3.1). By separately estimating depth on foreground and background, the proposed stereo matching solution will be also robust to depth discontinuities and misalignments of the flash and noflash images. On the other hand, the estimated depth can be integrated with the flash cue to overcome the limitations of FlashCut (Sun et al. 2007). These two problems are integrated in a uniform probabilistic framework and solved using Expectation-Maximization (EM) (in Sect. 2.3).

2.2 Model Initialization

Our system uses two stereo image pairs captured with and without flash. Let $d(p)$ denote the disparity at pixel p in the left/reference image F_l and $F_l(p)$ denote the color/intensity at p. A general assumption in stereo matching is that pixel p has the same color/intensity value or related transformed pattern (e.g., Census transform Zabih and Woodfill 1994) as the corresponding pixel in right/corresponding image F_r:

$$F_l(p) = F_r(p - d(p)). \tag{1}$$

Assuming that each scene contains a separable foreground and background layer and let $\alpha_l(p)$ and $\alpha_r(p-d(p))$ denote the foreground probability at pixel p in F_l and the corresponding pixel $p - d(p)$ in F_r, respectively. Some of the background pixels will be occluded by the foreground. This paper addresses this problem by performing stereo matching on foreground and background layer separately. Let F^f denote foreground layer of image F and F^b denote background layer of image F. From Eq. (1) we have:

$$\alpha_l(p)F_l^f(p) = \alpha_r(p + d^f(p))F_r^f(p + d^f(p)); \tag{2}$$

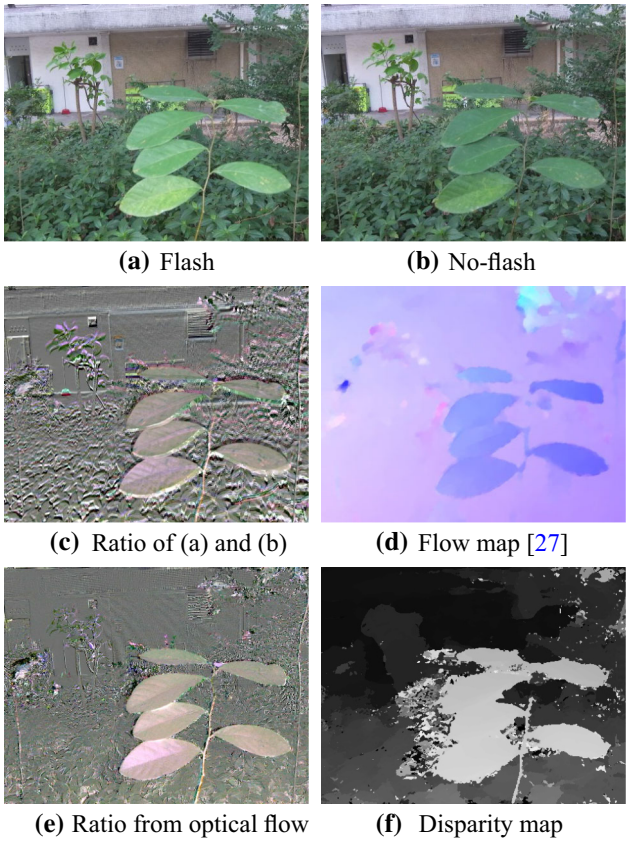
$$(1 - \alpha_l(p))F_l^b(p) = (1 - \alpha_r(p + d^b(p)))F_r^b(p + d^b(p)), \tag{3}$$

where $d^f(p), d^b(p)$ denote the disparities for latent pixels in foreground and background respectively. This paper aims at simultaneously estimating disparity d_p and foreground probability $\alpha_l(p)$ and $\alpha_r(p-d(p))$ using Expectation-Maximum (**EM**).

Before EM optimization (which is detailed in Sect. 2.3), all the unknown variables in the model will be initialized. The initial disparity map d can be estimated by a typical edge-preserving stereo matching algorithm (e.g., CostFilter Hosni et al. 2013) and the initial foreground/background layer F^f and F^b can be extracted from FlashCut (Sun et al. 2007).

Unlike FlashCut (Sun et al. 2007), we initialize the foreground probability α_l using motion compensated ratio map between flash left image and no-flash left image. Let Δ_p denote the motion between the flash and no-flash images obtained from optical flow (Weinzaepfel et al. 2013; Sun et al. Jan. 2014) or descriptor-based dense correspondence (Liu et al. 2008; Yang et al. 2014). The ratio at pixel p is computed as follows:

$$R(p)^{\Delta(p)} = \log \frac{F(p) + \epsilon}{N(p + \Delta(p)) + \epsilon}, \tag{4}$$



(a) Flash **(b)** No-flash

(c) Ratio of (a) and (b) **(d)** Flow map [27]

(e) Ratio from optical flow **(f)** Disparity map

Fig. 3 Direct integration with optical flow estimation is not practical as shown in the derived ratio map in **e** and the estimated disparity map in **f**

where, ϵ is a small value to avoid division by zero and F_p and $N_{p+\Delta(p)}$ are the pixel values of the flash and no-flash images at pixel p and $p+\Delta(p)$, respectively. Figure 3d presents the flow map computed using the state-of-the-art DeepFlow algorithm (Weinzaepfel et al. 2013) from Fig. 3a, b. Note that optical flow estimation cannot provide a reliable motion estimate at pixel level on real-world images, especially due to the significant brightness changes between the flash and no-flash image. As can be seen in Fig. 3e, the ratio map obtained from motion is very noisy in background. The disparity map obtained based on this ratio map in Fig. 3f shows that direct combination of optical flow estimation and the stereo method in (Zhou et al. 2012) is sensitive to misalignments of flash and no-flash images even for small movements presented in Fig. 3a, b.

This paper uses the motion compensated ratio map (in Fig. 3e) to estimate the initial foreground probability. First, the ratio value of each pixel is normalized to be an integer within (0, 255). After that, the ratio histogram is constructed. In general, a pixel with a large ratio value $R(p)^{\Delta(p)}$ is likely to be assigned into foreground layer and has large foreground probability. We model those pixels that take large ratio values in ratio histogram as a Gaussian distribution $\mathcal{N}(R(p)|\mu,\sigma^2)$ with mean μ and variance σ^2. Then the initial foreground probability of pixel p is further formulated as:

$$\alpha_l(p) = exp\left(-\sigma_i(R(p)-\mu)^2\right), \tag{5}$$

where $\sigma_i = ln2/(3\sigma)^2$. The initial foreground probability of flash right image $\alpha_r(p-d(p))$ can be estimated in the same manner.

2.3 Expectation-Maximization Optimization

In this section, we describe an EM optimization framework to estimate the parameters α_l, α_r with hidden data d^f, d^b. We want to maximize the log-likelihood $\log \sum_{d_i^f, d_j^b} P(d_i^f, d_j^b, \alpha_l, \alpha_r, F_l, F_r)$. Let L denote the disparity search range, we assume that both d_i^f and d_j^b range from 0 to $L-1$. With the estimated parameters α_l and α_r, we can further improve the accuracy of the foreground/background extraction (in Sect. 2.4) and disparity estimation on foreground and background layers (in Sect. 2.5).

2.3.1 E-step

In this step, given the foreground probabilities (α_l, α_r) of both left and right flash images, we compute the expectation of d^f and d^b which are disparities on the foreground and background layer, respectively. Assuming that the distributions of d^f and d^b are statistically independent, the expectation of $d^f(p)$ and $d^b(p)$ at pixel p can be formulated as:

$$\begin{aligned} &E(P(d^f(p), d^b(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r)) \\ &= E(P(d^f(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r) \\ &\quad \cdot P(d^b(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r)) \\ &= E(P(d^f(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r)) \\ &\quad \cdot E(P(d^b(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r)). \end{aligned} \tag{6}$$

The occluded regions will be located only on the background layer, and thus the expectation of foreground disparity d^f and background disparity d^b is computed separately as follows.

Expectation of Foreground Disparity According to Bayes' theorem, the conditional probabilities of foreground disparity d^f at pixel p is modeled as:

$$\begin{aligned} &P(d^f(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r) \\ &\propto P(\alpha_l^{(t)}, \alpha_r^{(t)}|d^f(p), F_l, F_r) \cdot P(d^f(p)|F_l, F_r), \end{aligned} \tag{7}$$

where $P(\cdot|d^f(p), F_l, F_r)$ is the conditional likelihood given d^f and $P(d^f(p)|F_l, F_r)$ is the prior probabilities of foreground disparity d^f.

Let likelihood $P(\alpha_l^{(t)}, \alpha_r^{(t)}|d^f(p), F_l, F_r)$ denote the similarity measurement between pixel p and its correspondence and both are on the foreground layer. According to Eq. 2, the matching cost at pixel p with disparity $d(p)$ is formulated as:

$$C^f(p) = \|\alpha_l(p)F_l^f(p) - \alpha_r(p - d^f(p))F_r^f(p - d^f(p))\|, \quad (8)$$

where $\|$ represents a distance measurement (e.g., L1 distance). We next model foreground disparity likelihood $P(\alpha_l^{(t)}, \alpha_r^{(t)}|d^f(p), F_l, F_r)$ as a Gaussian distribution $\mathcal{N}(C^f(p)|0, \sigma_{cf}^2)$ with mean 0 and variance σ_{cf}^2:

$$P(\alpha_l^{(t)}, \alpha_r^{(t)}|d^f(p), F_l, F_r) = \frac{1}{\sqrt{2}\sigma_{cf}} exp\left(-\frac{C^f(p)^2}{2\sigma_{cf}^2}\right). \quad (9)$$

σ_{cf} is a constant and set to 6 in our experiments. To suppress noise, an edge-preserving image filtering operation (*e.g.*, bilateral filtering (Tomasi and Manduchi 1998; Yang et al. 2009; Yang 2012), guided filtering (He et al. 2013) will be used to smooth $C^f(p)$ before computing the foreground likelihood in Eq. 9.

To model the foreground disparity distribution, we first generate a disparity histogram (as shown in Fig. 4) from disparity map computed in the previous iteration t or the initial disparity map for the first iteration. We model the disparity histogram with two Gaussian distributions: one models foreground disparities and the other models background disparities. Let $\mathcal{N}(d^f|\mu_{df}, \sigma_{df}^2)$ denote the foreground disparity distribution with mean μ_{df} and variance σ_{df}^2, and $\mathcal{N}(d^b|\mu_{db}, \sigma_{db}^2)$ denote the background disparity distribution with mean μ_{db} and variance σ_{db}^2. The prior probabilities of foreground disparity d^f is formulated as follows:

$$P(d^f(p)|F_l, F_r) = \frac{1}{\sqrt{2}\sigma_{df}} exp\left(-\frac{(d^f(p) - \mu_{df})^2}{2\sigma_{df}^2}\right). \quad (10)$$

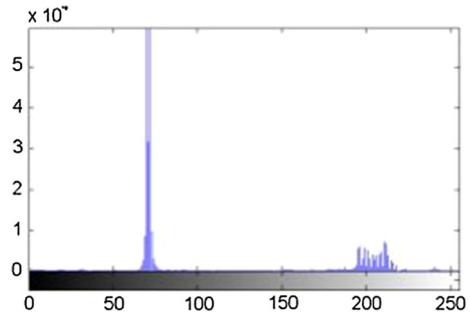


Fig. 4 A histogram of disparity map in Fig. 2e

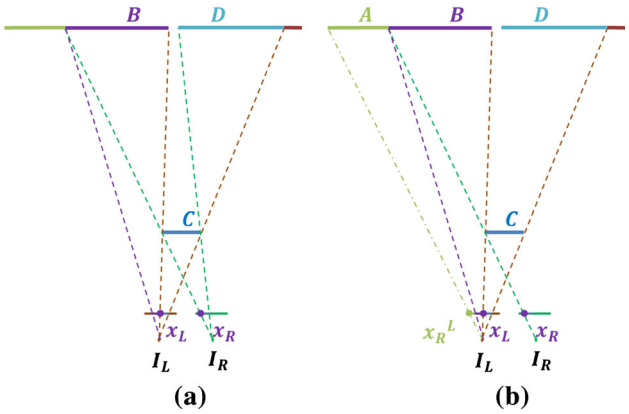


Fig. 5 Occlusion Detection

According to Eqs. (9) and (10), the expectation of foreground disparity $d^f(p)$ at pixel p is written as:

$$E(P(d^f(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r)) = \frac{P(d^f(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r)}{\sum_{d_i^f(p)} P(d_i^f(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r)}, \quad (11)$$

where, $d_i^f(p) \in \{0, 1, \cdots, L-1\}$.

Expectation of Background Disparity Similarly, the conditional probabilities of background disparity d^b at pixel p can be modeled as:

$$P(d^b(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r) \propto P(\alpha_l^{(t)}, \alpha_r^{(t)}|d^b(p), F_l, F_r) \cdot P(d^b(p)|F_l, F_r), \quad (12)$$

where $P(\cdot|d^b(p), F_l, F_r)$ is the conditional likelihood given d^b and $P(d^b(p)|F_l, F_r)$ is the prior probabilities of background disparity d^b.

We define the likelihood $P(\alpha_l^{(t)}, \alpha_r^{(t)}|d^b(p), F_l, F_r)$ as the similarity measurement between the corresponding pixels on background layers. Similar to Eq. (8), the matching cost at pixel p on the background is modeled based on Eq. (3):

$$C^b(p) = \|(1 - \alpha_l(p)F_l^b(p)) - (1 - \alpha_r(p + d^f(p)))F_r^b(p + d^b(p))\|. \quad (13)$$

Differing from the foreground pixels, some pixels on background are occluded and the matching costs of these pixels are not reliable for the computation of the background likelihood $P(\cdot|d^b(p), F_l, F_r)$.

A new occlusion detection method is proposed based on extracted foreground/background layers. The occlusion problem in binocular stereo matching refers to the fact that some points in the background are visible to only one camera, due to the shielding from foreground. For instance, the purple line B in Fig. 5a can be seen by the left camera but is invisible to the right camera. The projection of line B on

the left camera is thus the occlusion region. The occlusion information is encoded in the two known foreground layers. Let x_L denote the left intersection of a scanline and the foreground in the left image and x_R denote the corresponding intersection of the same scanline and the foreground in the right image as shown in Fig. 5a. x_L and x_R can be directly obtained from the foreground layer. Under the assumption of a rectified stereo pair, it is clear that the occlusion to the left of x_L is the purple line B in Fig. 5 a. Let x_R^L denote a pixel in the left image that has the same coordinates as x_R in the right image as shown in Fig. 5b), then scanline segment between pixel x_R^L and x_L will correspond to the combination of line segment A and B, which is obviously larger than or equal to the occlusion. It is equal to the occlusion only when line A is infinitely far way from the left camera. In practice, most flash and no-flash image pairs based computer vision and computer graphics tasks assume that the background is sufficient far away from the foreground, it is safe to assume that the region between x_R^L and x_L is a good approximation of the occlusion. In practice, it will never be smaller than the ground-truth occlusion region, which means that we can use the foreground layer to exclude all pixels reside in the occluded regions, and very likely, a small amount of pixels on left of the occlusion will be also excluded. Nevertheless, the removal of these small amount of non-occluded pixels will not deteriorate the performance much as has been demonstrated by many weighted median filtering based disparity refinement approaches (Ma et al. 2013). Figure 6e presents the occluded areas detected by two extracted foreground layers presented in Fig. 6c, d. Note that the estimated occluded areas are marked in white.

Let $O(p)$ denote the occlusion mask value (1 for occluded areas and 0 for unoccluded areas) at pixel p. And the matching cost of pixel p is rewritten as:

$$\tilde{C}^b(p) = (1 - O(p))C^b(p) + O(p)C_o^b(p), \tag{14}$$

where $C^b(p)$ is the original matching cost from Eq. (13), and $C_o^b(p)$ denote the matching cost in occluded areas. To compute $C_o^b(p)$, we first fill matching cost in detected occlusion regions with 0, and then propagate the matching costs from the non-occluded neighborhood into occlusion regions, using edge-preserving image filtering techniques (Tomasi and Manduchi 1998; Yang et al. 2009; Yang 2012; Ma et al. 2013; He et al. 2013).

According to $\tilde{C}^b(p)$ in Eq. (14) , the background disparity likelihood $P(\alpha_l^{(t)}, \alpha_r^{(t)}|d^b(p), F_l, F_r)$ is modeled as a Gaussian distribution $\mathcal{N}(C^f(p)|0, \sigma_{cf}^2)$ with mean 0 and variance σ_{cb}^2:

$$P(\alpha_l^{(t)}, \alpha_r^{(t)}|d^b(p), F_l, F_r) = \frac{1}{\sqrt{2}\sigma_{cb}} exp\left(-\frac{\tilde{C}^b(p)^2}{2\sigma_{cb}^2}\right). \tag{15}$$

In our experiments, σ_{cb} is set as 6.

Similar to Eq. (10), the prior probabilities of background disparity d^b can be modeled as a Gaussian Distribution given an existing disparity map:

$$P(d^b(p)|F_l, F_r) = \frac{1}{\sqrt{2}\sigma_{db}} exp\left(-\frac{(d^b(p) - \mu_{db})^2}{2\sigma_{db}^2}\right), \tag{16}$$

where μ_{db} denote mean and σ_{db}^2 denote variance.

From Eqs. (15) and (16), the expectation of the background disparity at pixel p can be formulated as follows:

$$E(P(d^b(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r)) = \frac{P(d^b(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r)}{\sum_{d_i^b(p)} P(d_i^b(p)|\alpha_l^{(t)}, \alpha_r^{(t)}, F_l, F_r)}, \tag{17}$$

where $d_i^b(p) \in \{0, 1, \cdots, L-1\}$.

2.3.2 M-step

In this section, we maximize the expected log-likelihood with respect to the parameters α_l, α_r given F_l, F_r. Let $\mathcal{X} = \{\alpha_l, \alpha_r\}$ denote the parameter set, $\mathcal{D} = \{d_i^f, d_j^b\}$ denote the hidden data set and $\mathcal{O} = \{F_l, F_r\}$ denote the observation, then:

(a) Left image **(b)** Right image **(c)** Left foreground **(d)** Right foreground **(e)** Occlusion

Fig. 6 Occlusion detection. **a** and **b** are *left* and *right* color images, the white regions in **c** and **d** are the extracted left and right foreground layers, and the white region in **e** is the estimated occluded regions

$$\mathcal{X}^{(t+1)} \propto \arg\max_{\mathcal{X}} \sum_{\mathcal{D}\in\mathcal{S}} P(\mathcal{D}|\mathcal{X}^{(t)}, \mathcal{O}) \log P(\mathcal{X}, \mathcal{D}, \mathcal{O})$$
$$= \arg\max_{\mathcal{X}} \sum_{\mathcal{D}\in\mathcal{S}} P(\mathcal{D}|\mathcal{X}^{(t)}, \mathcal{O}) \log(P(\mathcal{D}, \mathcal{O}|\mathcal{X})P(\mathcal{X})) \tag{18}$$

where $\mathcal{S}$ is the disparity space which contains all combinations of disparity of foreground with that of background. Let $\mathcal{L}(\cdot)$ denote $\log P(\cdot)$,

$$\begin{aligned}
&\mathcal{L}(\mathcal{D}, \mathcal{O}|\mathcal{X}) \\
&\quad = \log P(\mathcal{D}, \mathcal{O}|\mathcal{X}) \\
&\quad = -\sum_{p\in F_l} \frac{(\alpha_l(p) - \alpha_r(p - d(p)))^2}{2\sigma_\alpha^2}; \qquad (19) \\
&\mathcal{L}(\mathcal{X}) \\
&\quad = \log P(\mathcal{X}) \propto \log P(\alpha_l) + \log P(\alpha_r) \\
&\quad = -\sum_{p\in F_l} \frac{(\alpha_l(p) - \mu_{\alpha_l})^2}{2\sigma_{\alpha_l}^2} \\
&\qquad -\sum_{p\in F_l} \frac{(\alpha_r(p - d(p)) - \mu_{\alpha_r})^2}{2\sigma_{\alpha_r}^2}. \qquad (20)
\end{aligned}$$

Eq. (19) models the distribution of foreground probability difference between left and right image as Gaussian distribution with mean 0 and variance σ_α^2 (which is set to 0.06 in our experiments) and Eq. (20) models the distribution of α_l and α_r as Gaussian distributions with mean μ_{α_l} and μ_{α_r} and variance $\sigma_{\alpha_l}^2$ and $\sigma_{\alpha_r}^2$, respectively.

Equation (18) can be solved with Eqs. (19) and (20) by minimizing $-\sum_{\mathcal{D}\in\mathcal{S}} P(\mathcal{D}|\mathcal{X}^{(t)}, \mathcal{O})(\mathcal{L}(\mathcal{D}, \mathcal{O}|\mathcal{X}) + \mathcal{L}(\mathcal{X}))$, and $\mathcal{X}^*$ denote the final parameter set.

2.4 Improving Foreground/background Extraction

After EM optimization, the best parameters α_l, α_r are estimated. Similar to FlashCut (Sun et al. 2007), we formulate foreground/background segmentation of left image F_l as a binary labeling problem and model our Markov Random Field (MRF):

$$E(B) = \sum_p E_d(b(p)) + \lambda \sum_{p,q} E_s(b(p), b(q)), \tag{21}$$

where B is the binary foreground mask, b_p is the mask value (1 for foreground and 0 for background) at pixel p, and λ is a scaling factor is set to 12 in our experiments. E_s is the smoothness term which penalizes the different labeling ($b(p)$,$b(q)$) for two adjacent pixels (p, q) in textureless areas. Similar to FlashCut (Sun et al. 2007), it is formulated as:

$$E_s(b(p), b(q)) = |b(p) - b(q)| \cdot \exp\left(-\beta\|F_l(p) - F_l(q))\|^2\right), \tag{22}$$

where $\beta = (2\langle\|F_l(p) - F_l(q)\|^2\rangle)^{-1}$ (Blake et al. 2004) and $\langle\cdot\rangle$ denotes the expectation. Note that the MRF energy in Eq. (21) can be optimized using Graph Cuts (Boykov et al. 2001) or Belief Propagation (Sun et al. 2003).

For the data term E_d, we model it based on the the foreground and color likelihood:

$$E_d(b(p)) = E_r(b(p)) + \eta E_f(b(p)) + \zeta E_c(b(p)), \tag{23}$$

where η and ζ are two factors and both are set to 0.1. E_r follows the flash cue used in FlashCut (Sun et al. 2007) which assumes that foreground pixels have lager flash differences. E_f tends to label the pixels that have large foreground probabilities (α_l that have been estimated from EM optimization in Sects. 2.3.1 and 2.3.2) as foreground. Based on the foreground probabilities, E_c models the foreground and background color likelihoods using Gaussian Mixture Models (GMMs).

Flash Term Similar to FlashCut (Sun et al. 2007), we model the global influence of flash on foreground with histogram based flash ratio:

$$r(p)^f = \max\left\{\frac{h^f_{ind_p} - h^{nf}_{ind_p}}{h^f_{ind_p}}, 0\right\}, \tag{24}$$

where ind_p is the bin index at pixel p, $h^f_{ind_p}$ and $h^{nf}_{ind_p}$ are bin values in the histograms of the flash image and no-flash image respectively. This flash ratio tends to assign larger values to the pixels which are brightened by flash and we are more likely to label these pixels as foreground. More details can be found in Sun et al. (2007). The energy of flash term is thus modeled as:

$$E_r(b(p)) = \begin{cases} 2(\max\{r(p)^f, 0.2\} - 0.2), & b(p) = 0; \\ 2(0.2 - \min\{r(p)^f, 0.2\}), & b(p) = 1. \end{cases} \tag{25}$$

Foreground Term The energy of foreground term is modeled based on the foreground probabilities α_l:

$$E_f(b(p)) = \begin{cases} 2\max\{\alpha_l(p), 0.6\} - 1.2, & b(p) = 0 \\ 0.8 - 2\min\{\alpha_l(p), 0.4\}, & b(p) = 1 \end{cases} \tag{26}$$

This term gives penalties to pixels which have been assigned into background layer but has a higher foreground probabilities (>0.6). In contrast, pixels with lower foreground probabilities (<0.4) are more likely to be assigned to background layer. Unlike FlashCut (Sun et al. 2007), the depth cue

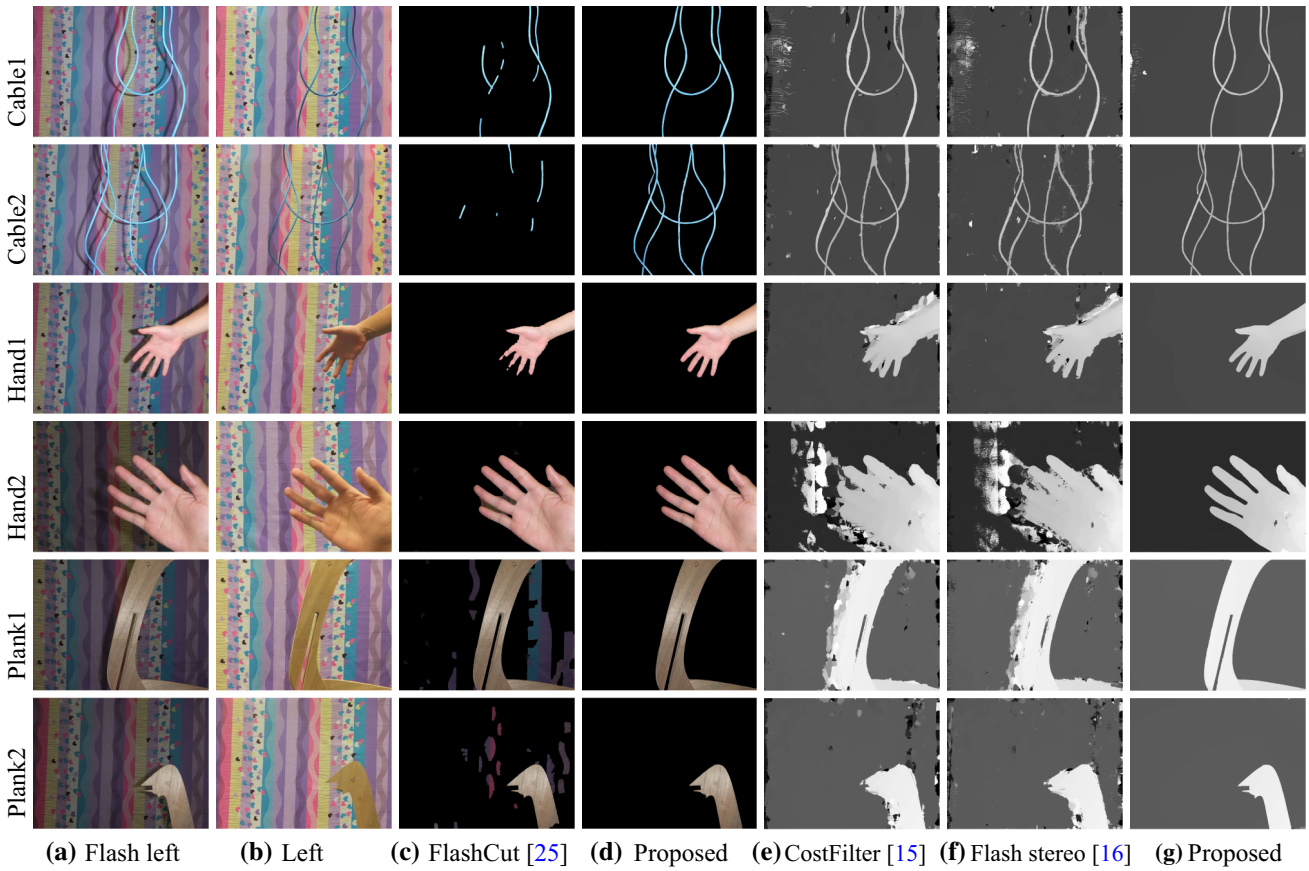


Fig. 7 Evaluation on data sets captured in indoor environments. **a**, **b** are the flash and no-flash images captured from the left camera, and there are some misalignments between them due to object movement or camera shake. **b**, **c** are the foreground layers extracted by FlashCut and the proposed system. **e**–**g** are the disparity maps computed from CostFilter, Flash stereo and the proposed system, respectively. Unlike the state-of-the-art, the proposed stereo system is robust to occlusions, depth discontinuities and misalignments

in this term is introduced by using optimal parameters (foreground probabilities α_l) estimated from EM optimization in Sects. 2.3.1 and 2.3.2.

Color Term Based on foreground probabilities α_l, the foreground/background color likelihood from depth is modeled as Gaussian Mixture Models (GMMs) (Blake et al. 2004):

$$p_c(F_l(p)|b(p)=1) = \sum_{i=1}^{K} \phi_i^f \mathcal{N}\left(F_l(p)|\mu_i^f, \Sigma_i^f\right), \quad (27)$$

where $F_l(p)$ is the color value at pixel p and $K = 10$ is the number of components. The i^{th} component of this foreground color GMMs is characterized by normal distributions with the weight ϕ_i^f, mean μ_i^f and covariance matrix Σ_i^f. The pixels with higher depth foreground probabilities (>0.6) will be used to train the foreground GMMs while pixels with the lower depth foreground probabilities (<0.4) are gathered to train the background GMMs. The color term are formulated as:

$$E_c(b(p)) = \begin{cases} -\log(p_c(F_l(p)|b(p)=1)), & b(p)=1; \\ -\log(p_c(F_l(p)|b(p)=0)), & b(p)=0. \end{cases} \quad (28)$$

2.5 Improving Disparity Estimation

The EM optimization gives reliable matching cost. This section introduces an adaptive cost aggregation method in this section to further improve disparity map. It is robust to occlusion, depth discontinuities and misalignments. With the binary foreground mask B obtained from Sect. 2.4, the foreground layer F_l^f and background layer F_l^b of flash left image are updated:

$$F_l^f(p) = b(p) \cdot F_l(p), \quad (29)$$

$$F_l^b(p) = (1 - b(p)) \cdot F_l(p). \quad (30)$$

These two color layers are used as guidance of the popular guided filtering technique (He et al. 2013) for efficient cost aggregation. We also tested other fast edge-preserving filters including the domain transform filter (Gastal and Oliveira

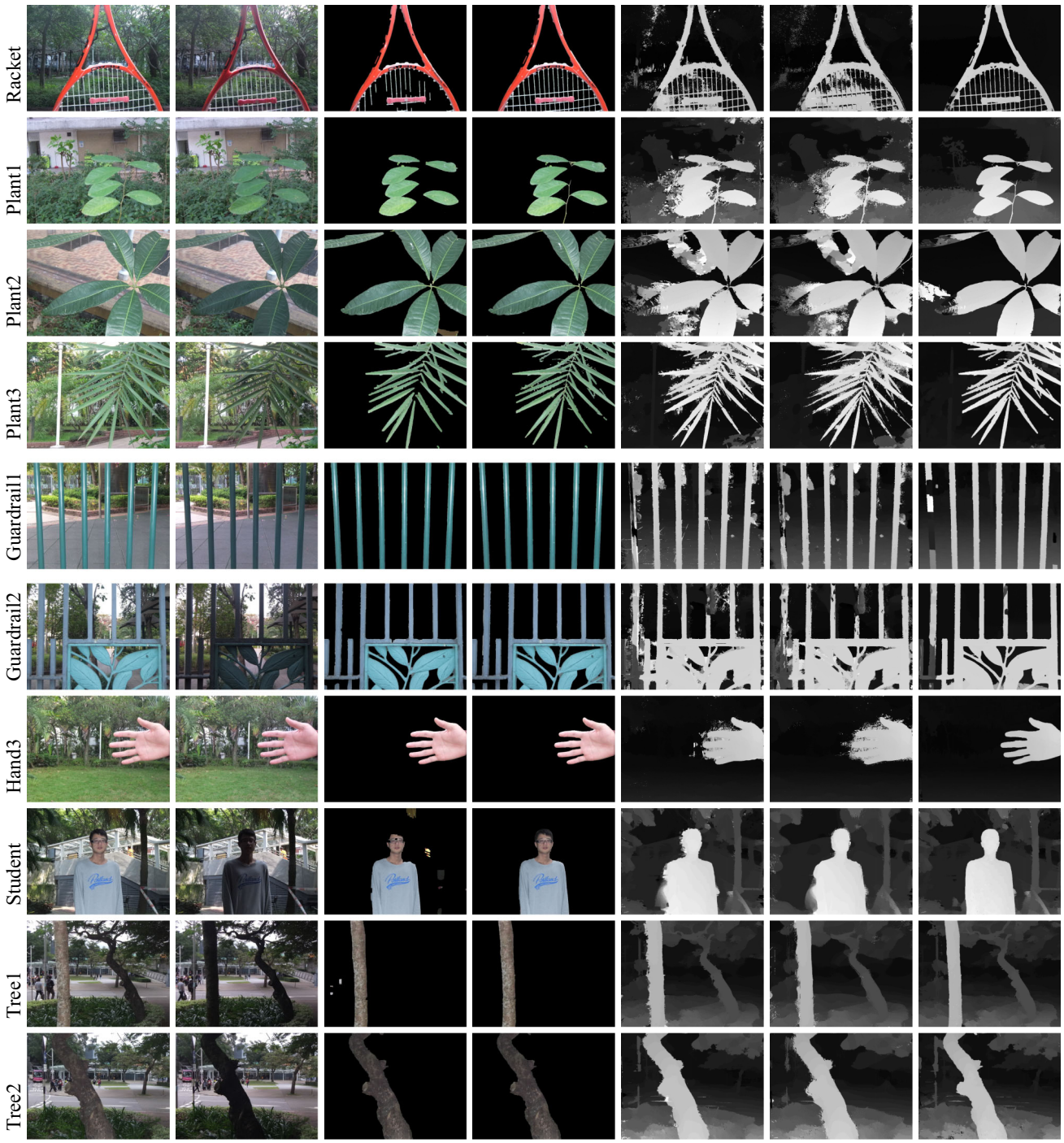


Fig. 8 Evaluation on data sets captured in outdoor environments. **a**, **b** are the flash and no-flash images captured from the left camera, and there are some misalignments between them due to object movement or camera shake. **b**, **c** are the foreground layers extracted by FlashCut and the proposed system. **e**–**g** are the disparity maps computed from CostFilter, Flash stereo and the proposed system, respectively. Unlike the state-of-the-art, the proposed stereo system is robust to occlusions, depth discontinuities and misalignments

2011), non-local filter (Yang 2012) and the recursive bilateral filter (Yang 2012). But according to our experiments, the performance of the guided image filter is the highest as it is difficult for the other fast filters to propagate reliable matching cost from non-occluded pixels to occluded pixels around texture regions.

The detailed aggregation using the two color guidance images F^f and F^b can be expressed as follows:

$$C_I^f(p) = \sum_{q \in \mathcal{N}_p} \omega_q^f \cdot b(q) \cdot C_{EM}^f(q), \tag{31}$$

$$C_I^b(p) = \sum_{q \in \mathcal{N}_p} \omega_q^b \cdot (1 - b(q)) \cdot C_{EM}^b(q), \tag{32}$$

where $\mathcal{N}_p$ is a local patch around pixel p, $C_{EM}^f(q)$ = -log $P(d^f(q)|\mathcal{X}^*, \mathcal{O})$ is the foreground matching cost at pixel q and $C_{EM}^b(q)$= $-\log P(d^b(q)|\mathcal{X}^*, \mathcal{O})$ is the background matching cost at pixel q. ω_q^f and ω_q^b are the supporting weights assigned to pixel q (by the adopted guided image filter). The final cost is simply a direct combination of the two aggregated costs at each pixel location:

$$C_I(p) = C_I^f(p) + C_I^b(p). \tag{33}$$

The benefit of this fusion is that it have the similar property of ratio image at depth discontinuities between foreground and background. The matching cost from foreground pixels will not be propagated into the background and thus will not deteriorate the detected occlusions. Meanwhile, the unreliable edges and ratio measures from motion compensation are disposed with the use of the clean color information.

After cost aggregation, a simple Winner-Takes-All (WTA) is performed to calculate disparity for each pixel.

3 Experimental Results

We evaluate our system on different flash/no-flash stereo image pairs captured by Fujifilm FinePix Real 3D W1 camera which is a standard consumer stereo camera. Each data set contains two stereo image pairs (with and without flash). There will be misalignments between every two flash and no-flash images. They are caused either by object movement or camera shake. The stereo images are rectified after camera calibration.

The proposed system is evaluated against the state-of-the-art techniques: (1) FlashCut foreground segmentation method (Sun et al. 2007) which uses a flash/no-flash image pair; (2) CostFilter stereo matching method (Hosni et al. 2013) which uses a stereo pair and adopts the guided image filter (He et al. 2013) to maintain the depth edges; and (3) Flash stereo (Zhou et al. 2012) which also uses a flash/no-flash stereo image pair.

According to our experiments, the proposed algorithm normally converges in only 2 iterations and thus we manually fix the number of iterations to be 2. A 3 × 3 Census transform (Zabih and Woodfill 1994) is used to compute the matching cost as it has been proven to be robust to outliers and radiometric differences in real environments (Hirschmuller

Table 2 Numerical evaluation of depth estimation on three indoor data sets

	Errors in all (%)			Errors in occluded areas (%)			Errors in discontinuities (%)			Errors in non-occluded areas and non-discontinuities (%)		
	Board	Plank3	Plank4	Board	Plank3	Plank4	Board	Plank3	Plank4	Board	Plank3	Plank4
CostFilter (Hosni et al. 2013)	5.04	4.73	4.07	60.04	55.96	62.65	19.36	17.55	19.22	0.99	1.51	0.92
Flash stereo (Zhou et al. 2012)	4.95	4.51	3.63	62.16	54.71	59.15	20.40	18.57	19.26	0.58	0.96	0.38
Proposed	1.66	2.34	1.28	3.62	8.96	3.71	7.30	10.97	6.89	0.06	0.16	0.11

The "Errors in all (%)" columns are percentages of bad pixels in the whole reference images; "Errors in occluded areas (%)" columns are percentages of bad pixels in occluded areas; "Errors in discontinuities (%)" columns are percentages of bad pixels in discontinuities; "Errors in non-occluded areas and non-discontinuities (%)" columns are percentages of bad pixels outside occluded areas and discontinuities. Note that proposed system (shown in the bottom row) consistently outperforms the state-of-the-art

Table 3 Numerical comparisons of foreground extraction

	Errors in all(%)			Errors in discontinuities (%)			Errors in non-discontinuities (%)		
	Board	Plank3	Plank4	Board	Plank3	Plank4	Board	Plank3	Plank4
Flash cut (Sun et al. 2007)	3.61	7.90	4.18	15.63	28.80	21.89	0.20	2.64	0.49
Proposed	1.65	2.31	1.24	7.28	10.83	6.89	0.06	0.16	0.06

The "Errors in all (%)" columns are percentages of bad pixels in the whole reference images; "Errors in discontinuities (%)" columns are percentages of bad pixels in discontinuities; "Errors in non-discontinuities (%)" columns are percentages of bad pixels outside discontinuities. Note that the reference foreground masks were extracted by manually thresholding on the (ground truth) disparity maps first. Then the accuracy were evaluated by counting the number of pixels with foreground binary values that differ from reference foreground masks. Proposed solution can produce more accurate foreground masks than those obtained from Flash cut (Sun et al. 2007)

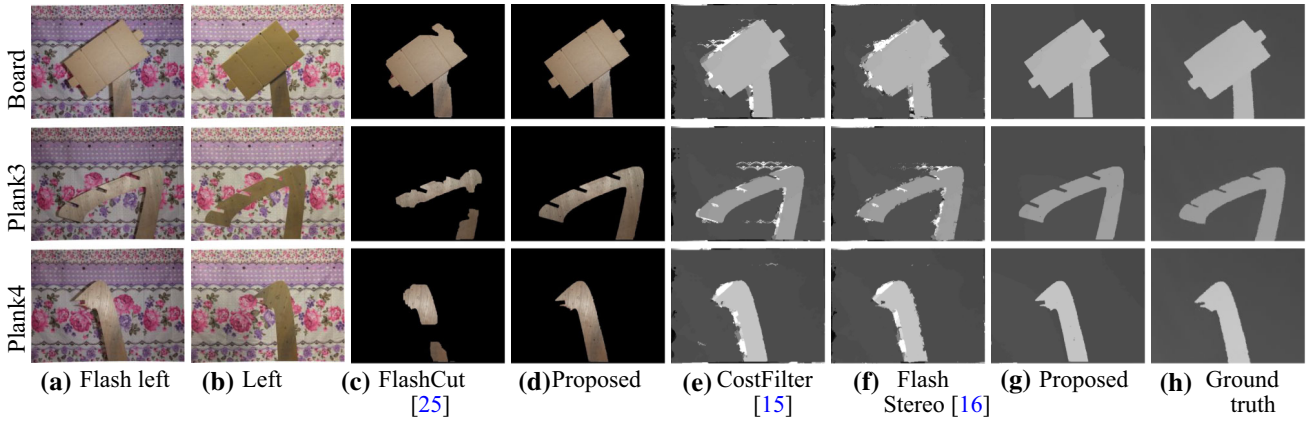


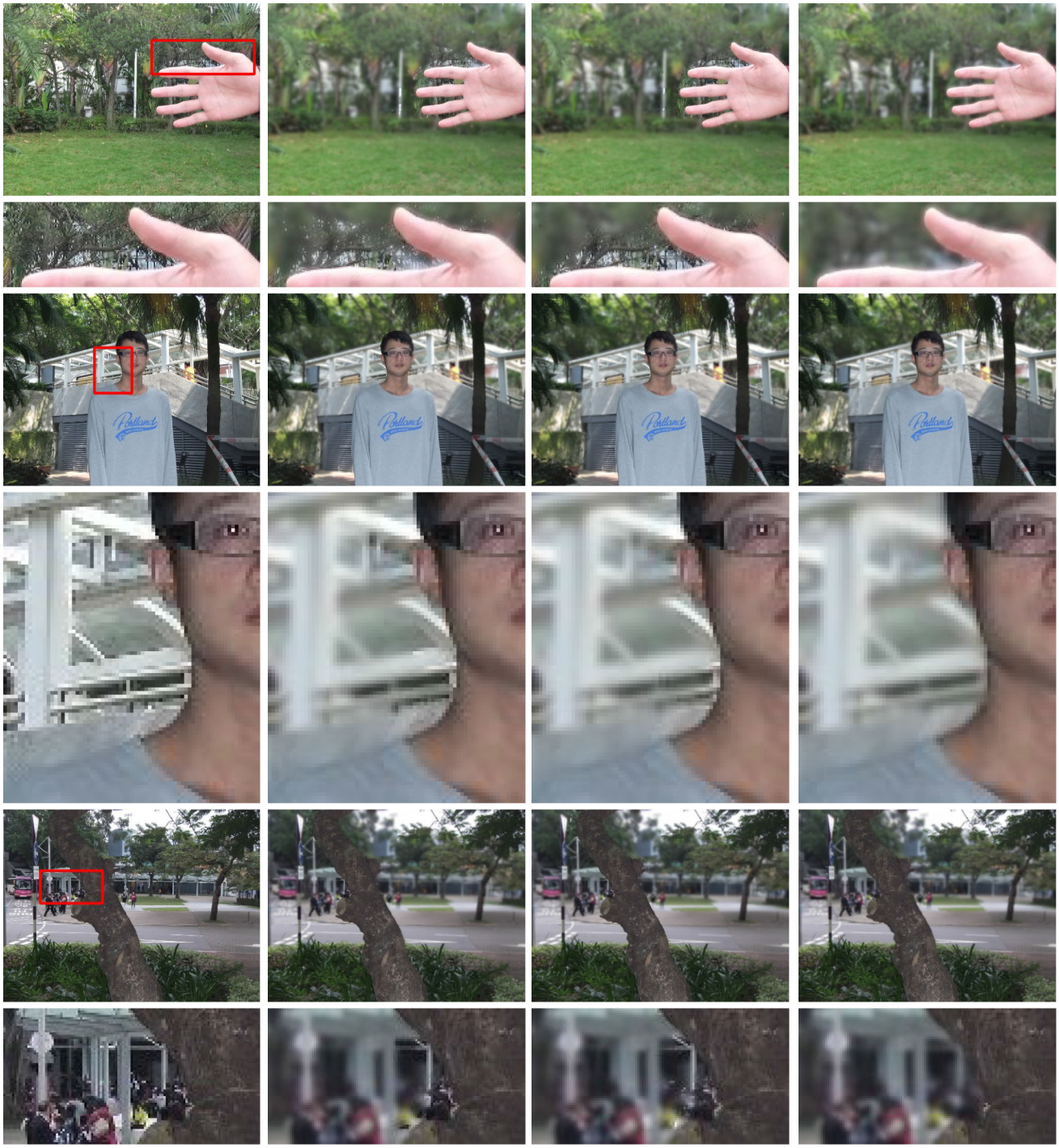
Fig. 9 Evaluation with the ground truth. **a**, **b** are the flash and no-flash images captured from the left camera, and there are some misalignments between them due to object movement or camera shake. **b**, **c** are the foreground layers extracted by FlashCut and the proposed system. **e**–**g** are the disparity maps computed from CostFilter, Flash stereo and the proposed system, respectively. **h** is the ground truth. As can be seen, the disparity maps obtained from the proposed stereo system appear to be quite close to the ground truth. The quantitative evaluation presented in Table 2 confirms the visual evaluation

and Scharstein 2009). To make sure that the comparison is fair, the guided image filter (He et al. 2013) is adopted in the cost aggregation step for all the stereo methods (although the original Flash Stereo Zhou et al. 2012 uses joint bilateral filter). The guided image filter parameters are set to $r = 15$ and $\epsilon = 8$. A winner-takes-all (WTA) selection will be performed after cost aggregation to compute the disparity values (without any post-processing step).

3.1 Visual Evaluation

Figures 7 and 8 present the visual comparisons on several data sets captured in the indoor and outdoor environments, respectively. As can be seen, FlashCut (Sun et al. 2007) is not suitable for indoor scenes in Fig. 7 as it will be hard to separate the foreground and background objects using flash. FlashCut (Sun et al. 2007) has problems with the thin-structured foreground objects as well. For instance, the blue cables in the "Cable1" data set in Fig. 7 and the white strings in the "Racket" data set and the plant stems in the "Plant1" data set in Fig. 8. It is difficult for Flash to brighten these objects as they are either in white/black color or have different surface orientations to flash near boundaries. In contrast, our segmentation results are more robust to these limitations thanks to the depth cue that compensates the flash limitations.

Figures 7 and 8 contain complex real-world scenes. Most of them are very challenging for stereo matching. Figures 7e and 8e present the disparity maps computed from CostFitler stereo method. We can see that this traditional method is vulnerable to occlusions and cannot successfully distinguish color edges from depth edges. There will be visible errors when the foreground is visually similar to the background (*e.g.*, leaves in "Plant1"). On the other hand, Flash stereo (Zhou et al. 2012) uses the ratio between flash and no-flash image pair which provides a more accurate guidance for matching cost aggregation than the original color information. The disparity maps obtained from Flash stereo (Zhou et al. 2012) are presented in Figs. 7f and 8f. Flash stereo assume no motion between the flash and no-flash stereo pair, and thus its performance is even lower than CostFilter in Figs. 7f and 8f due to inevitable motion in practice. Besides, both CostFilter and Flash stereo are vulnerable to occlusions. The disparity maps computed from the proposed system are presented in Figs. 7g and 8g. It is visually much more robust to depth discontinuities and occlusions when camera and/or object motion is presented.



(a) Original left image **(b)** CostFilter [15] **(c)** Flash stereo [16] **(d)** Proposed

Fig. 10 Bokeh effect. The proposed system is occlusion-aware which is essential for producing bokeh effect for a commercial stereo cameras. The state-of-the-art stereo algorithms can easily introduce noticeable artifacts around occlusions as shown in **b**, **c**. The proposed system can better suppress these artifacts as demonstrated in **d**

3.2 Numerical Evaluation

We further evaluate the performance of our system on several indoor data sets [1] with ground-truth disparity maps generated from Kinect sensor and Kinect fusion technique (Izadi et al. 2011). The reconstruction accuracy is measured by the percentage of bad pixels in both the whole reference image and the occlusion. Same as the Middlebury benchmark, the disparity error threshold is set to 1 and a pixel is treated a bad pixel if the difference between the estimated disparity and the ground truth is larger than the error threshold.

Table 2 presents the detailed numerical results. As can be seen, the performance of proposed system is much higher CostFilter and Flash Stereo, especially around occlusions and depth discontinuities. Table 2 also shows that although there are misalignments between the flash and no-flash images due to motion, Flash stereo still outperforms CostFilter when the whole image is considered. This is mainly because the background is highly-textured in color while the depth is almost constant. Using the color as guidance for cost aggregation is thus not suitable.

The numerical comparisons presented in Table 2 supports the visual comparisons in Figs. 7 and 8 that the proposed system is more robust to depth discontinuities and occlusions and can generate more accurate disparity maps. On the other hand, the proposed system can simultaneously produce a foreground mask which is more accurate than FlashCut (Sun et al. 2007). The detailed evaluation is presented Table 3. The three data sets used in this numerical evaluation is presented in Fig. 9. The foreground layer and disparity map estimated from different methods are presented from Fig. 9c–g, respectively.

3.3 Application to Bokeh Effect

The proposed system is occlusion-aware which is essential for producing bokeh effect for a commercial stereo cameras (e.g., Fujifilm FinePix Real 3D W3 camera or mobile phones like Huawei Honor 6 Plus). Figure 10 demonstrates that the proposed system outperforms traditional stereo algorithms for this application. It can better suppress visible artifacts thanks to the accurate depth estimates around depth edges. Figure 10 was produced based on the flash images of the *Hand3, Student and Tree2* data sets presented in Fig. 8a and the corresponding disparity maps in e–g.

3.4 Comparison of Computational Time

This section analyzes the computational cost of different stereo algorithms. Figure 11 compares the runtime of various algorithms on a typical data set (image size: 960×720; disparity range: $0 - 100$ pixels). The proposed algorithm takes over 140 s to process an flash/noflash stereo pair. It is relatively slower than the CostFilter (Hosni et al. 2013) ($\times 7$) and Flash stereo (Zhou et al. 2012) ($\times 3$). The experiments were conducted on a 1.7 GHz Intel Core i5 CPU. Parallel implementations can be used to accelerate the cost volume filtering operations (twice in initialization step and once in disparity map improving step after EM optimization in each iteration) adopted in proposed algorithm.

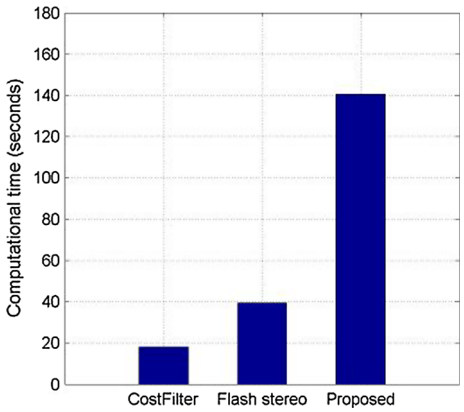


Fig. 11 Computational cost

4 Conclusion

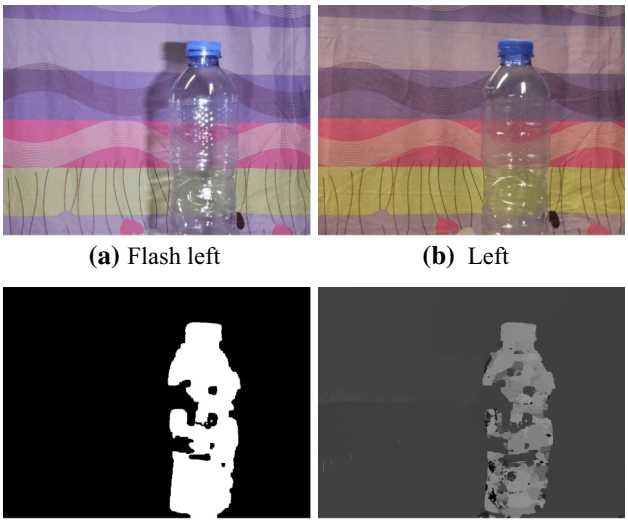
This paper presents a practical stereo system with a flash and no-flash stereo pair. Unlike the state-of-the-art Flash stereo algorithm (Zhou et al. 2012), it is robust to occlusions and inevitable misalignments between flash and no-flash image due to camera shake and scene motion. Additionally, it can simultaneously extract accurate foreground/background layer. Due to the integration of the depth information from stereo vision, it is much more robust to the traditional flash limitations and outperforms the state-of-the-art FlashCut foreground extraction technique (Sun et al. 2007).

Some scenes do not have clear foreground and background. For instance, slant surfaces (floor) in Fig. 12a, b. In this case, neither FlashCut (Sun et al. 2007) nor the proposed algorithm can extract accurate foreground layer. The difference between the flash and noflash image pair is closely related to the surface orientation and will be relatively small in this case. FlashCut (Sun et al. 2007) can completely fail on these scenes as shown in Fig. 12c. The proposed algorithm benefits from the fusion of depth cue and thus can produce a slightly more reliable foreground layer as shown in Fig. 12d. On the other hand, although accurate foreground/background layers are not reachable in this case, the adopted cost aggre-

[1] Most of the current commercial active sensors are not reliable under outdoor environment and thus only indoor environment was tested.

(a) Flash left **(b)** Left **(c)** FlashCut [25] **(d)** Proposed **(e)** CostFilter [15] **(f)** Flash Stereo [16] **(g)** Proposed

Fig. 12 Scenes without obvious foreground and background (e.g., slant surfaces). Neither FlashCut (Sun et al. 2007) nor the proposed algorithm can extract accurate foreground layer as shown in **c**, **d**. Nevertheless, all stereo algorithms are able to obtain correct disparity estimates as demonstrated in **e**–**g**



(a) Flash left **(b)** Left **(c)** Foreground **(d)** Disparity

Fig. 13 A failure case. Similar to the traditional stereo matching algorithms, the proposed stereo system is vulnerable to transparent and specular objects

gation method (proposed in Sect. 2.5) is still able to produce correct disparity map as can be seen from Fig. 12g.

However, the proposed stereo system shares some limitations with the traditional stereo matching algorithms. It is also vulnerable to transparent and specular objects and an example is presented in Fig. 13.

References

Bastanlar, Y., Temizel, A., Yardimci, Y., & Sturm, P. (2012). Multi-view structure-from-motion for hybrid camera scenarios. *Image and Vision Computing*, *30*(8), 557–572.

Blake, A., Rother, C., Brown, M., Perez, P., & Torr, P. (2004). Interactive image segmentation using an adaptive gmmrf model. In *ECCV* (pp. 428–441).

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *PAMI*, *23*(11), 1222–1239.

Chen, C., Lin, H., Yu, Z., Kang, S., & Yu, J. (2014). Light field stereo matching using bilateral statistics of surface cameras. In *CVPR*.

Gastal, E. S. L., & Oliveira, M. M. (2011). Domain transform for edge-aware image and video processing. *TOG*, *30*(4), 69:1–69:12.

He, K., Sun, J., & Tang, X. (2013). Guided image filtering. *PAMI*, *35*, 1397–1409.

Hirschmuller, H., & Scharstein, D. (2009). Evaluation of stereo matching costs on images with radiometric differences. *PAMI*, *31*(9), 1582–1599.

Hosni, A., Rhemann, C., Bleyer, M., Rother, C., & Gelautz, M. (2013). Fast cost-volume filtering for visual correspondence and beyond. *PAMI*, *35*, 504–511.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., & Fitzgibbon, A. (2011). Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *UIST* (pp. 559–568).

Kaehler, O., & Reid, I. (2013). Efficient 3d scene labeling using fields of trees. In *ICCV* (pp. 3064–3071).

Liu, C., Yuen, J., Torralba, A., Sivic, J., & Freeman, W. T. (2008). Sift flow: Dense correspondence across different scenes. In *ECCV* (pp. 28–42).

Ma, Z., He, K., Wei, Y., Sun, J., & Wu, E. (2013). Constant time weighted median filtering for stereo matching and beyond. In *ICCV*.

Murray, D., & Little, J. (2000). Using real-time stereo vision for mobile robot navigation. *Autonomous Robots*, *8*(2), 161–171.

Point-gray stereo camera. (2015). http://www.ptgrey.com//bumblebee2-firewire-stereo-vision-camera-systems.

Prisacariu, V., & Reid, I. (2012). 3d hand tracking for human computer interaction. *Image and Vision Computing*, *30*(3), 236–250.

Ren, C., Prisacariu, V., Murray, D., & Reid, I. (2013). Star3d: Simultaneous tracking and reconstruction of 3d objects using rgb-d data. In *ICCV* (pp. 1561–1568).

Riegl vz 1000 scanner. http://www.riegl.com/nc/products/terrestrial-scanning/produktdetail/product/scanner/27/.

Rothganger, F., Lazebnik, S., Schmid, C., & Ponce, J. (2006). 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *IJCV*, *66*(3), 231–259.

Scharstein, D., & Szeliski, R. Middlebury stereo evaluation. http://vision.middlebury.edu/stereo/eval/.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, *47*, 7–42.

Softkinetic depth sensor. (2015). http://www.softkinetic.com/Products/DepthSenseCameras.

Sun, J., Li, Y., Kang, S., & Shum, H. (2006). Flash matting. In *SIGGRAPH* (pp. 772–778).

Sun, J., Sun, J., Kang, S., Xu, Z., Tang, X., & Shum, H. (2007). Flash cut: Foreground extraction with flash and no-flash image pairs. In *CVPR*.

Sun, D., Roth, S., & Black, M. (2014). A quantitative analysis of current practices in optical flow estimation and the principles behind them. *IJCV*, *106*(2), 115–137.

Sun, J., Zheng, N., & Shum, H. Y. (2003). Stereo matching using belief propagation. *PAMI*, *25*(7), 787–800.

Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. In *ICCV* (pp. 839–846).

Weinzaepfel, P., Revaud, J., Harchaoui, Z., & Schmid, C. (Dec. 2013). Deepflow: Large displacement optical flow with deep matching. In *ICCV*, Sydney.

Xiong, W., Chung, H., & Jia, J. (2009). Fractional stereo matching using expectation-maximization. *PAMI*, *31*(3), 428–443.

Yang, Q. (2012). A non-local cost aggregation method for stereo matching. In *CVPR* (pp. 1402–1409).

Yang, Q. (2012). Recursive bilateral filtering. In *ECCV* (pp. 399–413).

Yang, H., Lin, W., & Lu, J. (2014). Daisy filter flow: A generalized discrete approach to dense correspondences. In *CVPR* (pp. 3406–3413).

Yang, Q., Tan, K.-H., & Ahuja, N. (2009). Real-time o(1) bilateral filtering. In *CVPR*.

Ye, J., Ji, Y., & Yu, J. (2013). A rotational stereo model based on xslit imaging. In *ICCV*.

Ye, J., Ji, Y., Li, F., & Yu, J. (2012). Angular domain reconstruction of dynamic 3d fluid surfaces. In *CVPR* (pp. 310–317).

Yoon, K.-J., & Kweon, I.-S. (2006). Adaptive support-weight approach for correspondence search. *PAMI*, *28*(4), 650–656.

Yu, Z., Guo, X., Ling, H., & Yu, J. (2013). Line assisted light field triangulation and stereo matching. In *ICCV*.

Zabih, R., & Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *ECCV*.

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE Multi-Media*, *19*(2), 4–12.

Zhou, C., Troccoli, A., & Pulli, K. (2012). Robust stereo with flash and no-flash image pairs. In *CVPR* (pp. 342–349).