Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

K-nearest neighborhood based integration of time-of-flight cameras and passive stereo for high-accuracy depth maps*

Li-wei LIU[†1,2], Yang LI[†1,2], Ming ZHANG[1,2], Liang-hao WANG[†‡1,2], Dong-xiao LI[1,2]

([1]*Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China*)

([2]*Zhejiang Provincial Key Laboratory of Information Network Technology, Hangzhou 310027, China*)

[†]E-mail: {lllw19870907, lychina, wanglianghao}@zju.edu.cn

Received July 15, 2013; Revision accepted Nov. 11, 2013; Crosschecked Feb. 19, 2014

Abstract: Both time-of-flight (ToF) cameras and passive stereo can provide the depth information for their corresponding captured real scenes, but they have innate limitations. ToF cameras and passive stereo are intrinsically complementary for certain tasks. It is desirable to appropriately leverage all the available information by ToF cameras and passive stereo. Although some fusion methods have been presented recently, they fail to consider ToF reliability detection and ToF based improvement of passive stereo. As a result, this study proposes an approach to integrating ToF cameras and passive stereo to obtain high-accuracy depth maps. The main contributions are: (1) An energy cost function is devised to use data from ToF cameras to boost the stereo matching of passive stereo; (2) A fusion method is used to combine the depth information from both ToF cameras and passive stereo to obtain high-accuracy depth maps. Experiments show that the proposed approach achieves improved results with high accuracy and robustness.

Key words: Depth map, Passive stereo, Time-of-flight camera, Fusion
doi:10.1631/jzus.C1300194 **Document code:** A **CLC number:** TP317.4

1 Introduction

Depth information is an important cue for the reconstruction of the corresponding captured real scene (Lee *et al.*, 2011). The obtained depth information can usually be used in many applications such as navigation and automobiles equipped with color cameras. Moreover, the depth information is also conducted to build digital textured models (Wang *et al.*, 2013) of scenes and objects to serve the purposes of documentation, planning, and visualization (May *et al.*, 2006; Attamimi *et al.*, 2010). Nowadays, the depth-image-based rendering approach (Buehler *et al.*, 2001) is attractive to applications in free viewpoint video and 3D television. Therefore, in a '2D+depth' 3DTV system, it is desirable to obtain the robust and accurate depth images of the relevant color images.

There exist a variety of technologies which are used to obtain depth information from a scene in the real world, such as passive stereo (e.g., stereo matching methods and photometric stereo), active stereo (e.g., laser range scanner), and time-of-flight (ToF) cameras (Xu *et al.*, 1998). Unfortunately, the aforementioned depth sensing methods are not perfect on their own individual performances. Stereo matching methods tend to perform poorly over the regions with textureless or repetitive patterns. Photometric stereo methods are prone to result in distortions. The efficiency of laser scanners is too low and cannot be used in a real-time setting. The resolution of the ToF cameras is very low and the captured data is often noisy. In this study, ToF cameras are cho-

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61072081 and 61271338), the National High-Tech R&D Program (863) of China (No. 2012AA011505), the National Science and Technology Major Project of the Ministry of Science and Technology of China (No. 2009ZX01033-001-007), the Key Science and Technology Innovation Team of Zhejiang Province (No. 2009R50003), and the China Postdoctoral Science Foundation (No. 2012T50545)

sen to capture real-time short-range depth images for further refinement.

ToF cameras can provide real-time independent 3D depth estimation at each valid pixel, and have been available only in recent years from companies such as Canesta (2006), PMD (2009), and 3DV (Z-cam, 2004). These devices use the principle of measuring the traveling time of actively transmitted light and do not depend on the color distribution of their captured scenes. ToF cameras do not suffer from the missing texture in the scenes or bad lighting conditions with less computational overload. However, ToF cameras usually have a low resolution (200×200 pixels) when compared with the full HD (1920×1080 pixels) television standard. Due to multiple interferences (e.g., noise and multiple reflections), the measured depth by ToF cameras may be incorrect and therefore a correction must be performed. To acquire more accurate depth information, ToF cameras provide not only real-time depth estimation but also the corresponding amplitude images of the scene. These amplitude images can be used to determine whether the estimated depth information of certain pixels is correct or not (Ringbeck and Hagebeuker, 2007). However, it is difficult for amplitude images to evaluate the estimated depth information of pixels in textureless regions. In fact, ToF cameras cannot provide accurate depth information for their captured scenes alone. Thus, it is imperative to introduce additional cues to improve the accuracy of depth information obtained by ToF cameras.

Passive stereo is the process of taking two or more images and estimating a 3D model of the scene by finding matching pixels in the images and converting their 2D positions into 3D depths. Nowadays, a vast majority of passive stereo matching research is devoted to obtaining depth information from images (De-Maeztu *et al.*, 2011; Yao *et al.*, 2012). Passive stereo matching usually performs well on textured regions. However, passive stereo typically experiences serious problems in the vicinity of 3D depth boundaries and fails within occluded areas and/or poorly textured regions (Scharstein and Szeliski, 2002a). Since ToF cameras and passive stereo are complementary to each other, this study attempts to improve the accuracy of depth images by combining the real-time depth estimation provided by the ToF sensors with the sophisticated algorithms of passive stereo matching.

This proposed approach is mainly divided into two steps. The motivation of the first step is to conduct both the depth estimation and the amplitude images from ToF cameras to boost the traditional passive stereo matching. To materialize the motivation, an energy cost function is devised to leverage ToF data (depth images and amplitude images) to guide the traditional stereo matching. This energy cost function is particularly appropriate for helping with stereo matching in traditional passive stereo over regions with textureless or repetitive patterns (e.g., the checkerboard) owing to the fact that ToF cameras will not be effected by texture in depth estimation. In the second step, the improved depth images from the passive stereo and the depth images from the ToF cameras are incorporated together to acquire the final high-accuracy depth images.

Although the proposed approach bears some resemblance with other approaches, it is worthwhile to highlight the main contributions of this study as follows: (1) An energy cost function is devised to boost the stereo matching of traditional passive stereo by the appropriate utilization of the data from ToF cameras; (2) A fusion mechanism is introduced to combine the depth information respectively by ToF cameras and passive stereo to obtain high-accuracy depth images.

2 Related works

One of the well established ToF depth enhancement approaches is Lindner *et al.* (2007)'s color image reference method. This method captures the ToF depth images and the referential high resolution color images with a combined vision system, consisting of a ToF camera and an RGB-camera. Under the assumption that the edges in color images are correlated with changes in the depth image, the depth data is projected onto the RGB image coordinates, then the edges and depth data are refined by the referential RGB images. However, the correlation assumption of the depth information and color distribution does not always hold since the surfaces might not necessarily be piecewise smooth. In fact, color images can easily produce a false depth discontinuity due to the variations of shading and illumination. Diebel and Thrun (2005) proposed another method for ToF depth refinement. In their work, a local cost function is defined based on Markov ran-

dom field (MRF) to enforce the spatial context for the neighboring pixels on depth images, and the best depth is chosen by the minimum cost of an MRF function. Unfortunately, the computational complexity of this method is very high.

Another portion of the existing ToF depth enhancement literature utilizes passive stereo algorithms to improve the accuracy of depth maps. Gudmundsson *et al.* (2008) proposed a ToF depth-based stereo matching method. This method uses two RGB cameras for stereo matching. In their method, ToF depth data is converted into parallax using parameters of the RGB cameras (e.g., focal length and optic center). This parallax is taken as an initial parallax for stereo matching refinement. Since the accuracy of depth maps by this method heavily relies on the accuracy of the initial ToF depth data, as introduced before, the depth information of noise and flying pixels will deteriorate the accuracy of the final depth images. Moreover, the errors in depth images due to multiple reflections are not discussed in this method. Zhu *et al.* (2008) proposed an approach to use ToF depth cameras together with stereo matching for enhancing depth maps and also computing alpha mattes. They formulate the problem of the depth reconstruction as an MRF to obtain the high-quality depth using global methods. One of the drawbacks of global methods is their computation overload is directly hindered by the number of pixels in the depth images. As a result, the parameter tuning to obtain a better result could thus be time consuming. To reduce the running times of global methods, this method also proposes an iterative approach as an approximation to belief propagation (BP). However, BP approximation will cause edge blur on the depth images. Improved depth results were demonstrated by Zhu *et al.* (2008), however the scenes in Zhu *et al.* (2008)'s experiments were captured in a controlled setting and therefore this approach is not suitable for complicated scenes.

There is also another proposal of ToF fusion recently presented by Gandhiy *et al.* (2012). This method is based on an efficient seed-growing algorithm which uses the ToF data projected onto the stereo image pair, and then these initial 'seeds' are propagated based on a Bayesian model. However, this method also suffers from the wrong depth estimation of ToF cameras, and therefore the wrong hypotheses of the ToF depth will be extended seriously in the final fusion depth map. Also, the experiments shown in Gandhiy *et al.* (2012) were captured in a controlled setting, in which the robustness of this method is questionable.

Since a matching cost is used to compute the disparities at each pixel and a suitable matching cost plays a fundamental role in passive stereo, we argue that the cues provided by ToF cameras can be used to define a more robust cost function to boost the passive stereo correspondence especially for the poorly-textured or occluded regions. Moreover, the depth map obtained by ToF cameras and stereo matching respectively is complementary to each other (i.e., low resolution versus high resolution and texture-insensitive versus texture-sensitive), and an appropriate fusion of the depth map by ToF cameras and stereo matching is attractive to acquire high-accuracy depth maps.

In a word, we adopt the data from ToF cameras to guide stereo matching of traditional passive stereo, and the stereo matched results in passive stereo and the depth information in ToF cameras are incorporated to obtain high-quality depth maps. The achieved results by the proposed approach are superior to these aforementioned global methods.

3 The proposed approach

In Section 2 we have discussed the limitations and challenges that the traditional stereo matching and ToF cameras encounter when they are individually used to obtain the accurate depth maps of corresponding scenes. Therefore, the proposed fusion approach contains two steps. First, since the stereo matching of traditional passive stereo over regions with textureless or repetitive patterns usually fails, the ToF cameras, which perform well over such regions, are used to refine the stereo matching of traditional passive stereo. Second, since the individually estimated depth information by stereo matching or ToF cameras will be incorrect with certain probabilities due to their innate limitations, the depth information by passive stereo correspondence and that by ToF cameras are complementary and are integrated to produce high-accuracy depth maps.

In the first step, different from other approaches, we focus on the utilization of a more robust cost function (i.e., an energy cost function) defined over each ToF depth pixel to guide the stereo matching of pas-

sive stereo. The underlying intuition of the devised energy cost function is that the closer the estimated depth is to the correct depth per pixel, the smaller the values of the energy cost function are. Otherwise, the values of the energy cost function will become larger. The energy cost function can improve the stereo matching results of regions with textureless or repetitive patterns by traditional passive stereo.

In the second step, a global regularization is devised to design a total optimization weighting function to produce the final high resolution depth map by the integration of ToF depth information, the stereo matching results, and the referential color images. The global regularization can guarantee that the acquired depth information for each pixel closely corresponds to the correct depth information.

To fuse the low-resolution estimated depth image of ToF cameras and the high-resolution estimated depth image of stereo matching, this approach warps the ToF depth map into the reference color view coordinates. This is done via a forward mapping from a low resolution one to a high resolution one. Fig. 1 gives an overview of this process. The details are explained in next sections.

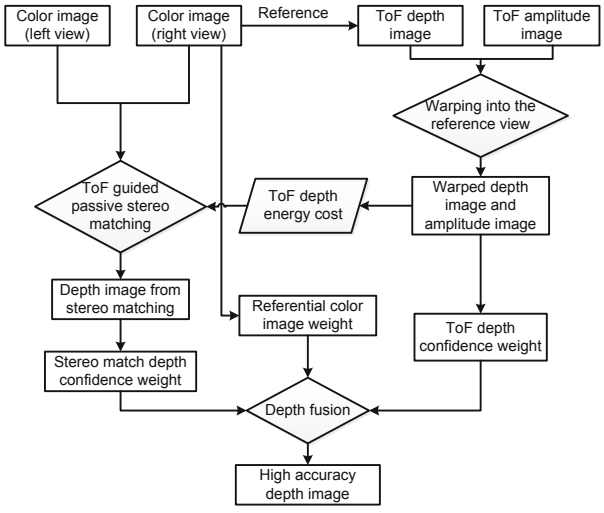

Fig. 1 Algorithmic framework of the proposed approach

4 ToF and RGB camera setup

Our lab combines a ToF camera with a pair of CCD HD color cameras to capture real-world scenes (Fig. 2). The ToF camera used in the capturing system is a PMD camcube3.0 (PMD, 2010), which can provide a depth map of 200×200 resolution with an operational range between 0.5 m and 7.5 m in real time. In the current setup, these two CCD HD color cameras are placed in parallel to each other. To facilitate stereo matching, this setup is designed to provide the pair of color images which have only horizontal parallax with each other.

Fig. 2 Camera setup of our capturing system

Due to radial lens distortion, the distortions for both color images and ToF depth images will be rectified in advance to prevent not only the false stereo matching between color images but also the false warping of the ToF depth images. The intrinsic and extrinsic parameters of three cameras, as well as their distortion coefficients, have been computed using computer vision library OpenCV (Opencv, 2012).

5 ToF guided passive stereo matching

5.1 Traditional stereo matching process

As discussed earlier, the data from ToF cameras will be used to boost the performance of traditional stereo matching. Here adaptive window based (Kanade and Okutomi, 1994; Yang, 2012) matching is used for stereo matching. The matching method will match small image windows centered at a given pixel, assuming that the visual characteristics are similar. During the stereo matching process, a pair of parallel color images captured by the two HD RGB cameras at the same time are used for stereo matching after both of them are rectified. Usually for each pixel (x_0, y_0) on the first (right) image I_r, stereo matching tends to find a disparity value m which will yield a correspondent image location

$[(x_0-m, y_0), (x_0+m, y_0)]$ on the second (left) image I_l. The procedure to compute disparity m usually requires three steps: local stereo matching, cost aggregation, and global optimization. It is a well known fact that the local stereo matching is the most important part.

For each pixel (x_0, y_0) on the right image I_r, we search for a certain horizontal neighborhood $[(x_0-m, y_0), (x_0+m, y_0)]$ of pixel (x_0, y_0) with the same coordinate on the left image I_l to calculate the matching cost. The original matching cost C_o is defined as the sum-of-absolute-distance (SAD) of two pixels in terms of their difference in RGB color spaces:

$$C_\mathrm{o}(x, y_0) = |I_\mathrm{l}(x, y_0) - I_\mathrm{r}(x_0, y_0)|, \quad (1)$$

where $(x, y_0) \in [(x_0-m, y_0), (x_0+m, y_0)]$. $[(x_0 - m, y_0), (x_0 + m, y_0)]$ is a searching range on the left image. Pixel $I_\mathrm{l}(x_\mathrm{min}, y_0)$ on the left image which has the minimum matching cost $C_\mathrm{o}(x_\mathrm{min}, y_0)$ compared with $I_\mathrm{r}(x_0, y_0)$ suggests that $I_\mathrm{l}(x_\mathrm{min}, y_0)$ is the correspondence pixel of $I_\mathrm{r}(x_0, y_0)$ in the right image. The x-axis coordinate difference of these two matched pixels shall closely correspond to the correct parallax $P_\mathrm{r}(x_0, y_0)$=$|x_\mathrm{min}-x_0|$ for pixel $I_\mathrm{r}(x_0, y_0)$. Then the depth map can be computed using the intrinsic parameters of the RGB cameras and the parallax map. The details of conversion between the depth map and parallax map will be discussed in Section 5.3.

However, due to the disturbance of noise, occluded parts, and poorly-textured regions, it is questionable in practice to pick up the true correspondence pixels from the right and left images. After the local stereo matching is done, the matching costs of pixels in the local window are aggregated (cost aggregation) to form a window based matching cost. In comparison to local stereo matching, more sophisticated solutions are offered in the form of global optimization. The global optimization is less sensitive to initial disparity values by local stereo matching and in principle more powerful, but usually requires considerably more memory storage and computational time, which makes a real-time application hardly conceivable. As a result, it is beneficial to introduce additional cues to efficiently improve the performance of stereo matching rather than global optimization.

5.2 Design of the ToF energy cost function

Since the pixels in textureless and repetitive patterns often have multiple minimum local matching cost values, it is difficult to determine the correct parallax through this cost C_o. So in our proposal, an energy cost function is devised to use ToF data (depth information and amplitude images of the corresponding scenes) to update the matching cost C_o. To prevent wrong depth estimation of the ToF camera, like the drawback of the method proposed by Gudmundsson *et al.* (2008), the reliability of ToF depth images will be detected in our approach. Then the reliable regions of ToF depth maps are used to guide the stereo matching of the passive stereo. The purpose of such attempts is to reduce the mismatch of stereo matching over regions with the textureless and repetitive patterns and also to acquire a better depth image.

ToF cameras in this study can provide two kinds of different images, namely ToF depth images and ToF amplitude images. Since these two kinds of images provide important cues to improve the performance of traditional stereo matching, the energy cost function is defined by these two kinds of images. The energy cost function consists of three terms: (1) an amplitude reliability weight, (2) a boundary reliability weight (these two weights are designed to find the reliable regions in ToF depth images), and (3) a depth difference penalty cost. We will clarify these three terms in detail as follows.

5.2.1 Amplitude reliability weight

Although ToF cameras can obtain the depth information over the regions with textureless or repetitive patterns, the estimated depth over low reflectivity objects in such regions is usually not reliable. Fortunately, given an amplitude image A which is corresponding to a ToF depth image from pixel to pixel, the value of pixel (x_0, y_0) in the amplitude image A, referred to as $A(x_0, y_0)$, can provide a cue to indicate whether the depth information by ToF cameras is reliable or not: the larger the value of $A(x_0, y_0)$ gets, the more reliable the depth information of the pixel is, due to its good reflectivity. According to such observations, the amplitude reliability weight is defined to denote the reliability of the estimated ToF depth for each pixel. The amplitude reliability weight $W_\mathrm{A}(x_0, y_0)$ of pixel (x_0, y_0) in the

amplitude image A is defined as follows:

$$W_{\mathrm{A}}(x_0,y_0)=\begin{cases}\frac{\lg A(x_0,y_0)+\delta}{\gamma}, & \text{if } A_{\min}\le A\le A_{\max},\\ 0, & \text{otherwise,}\end{cases} \tag{2}$$

where δ and γ are manually predefined parameters. Obviously, the amplitude reliability weight equals 0 if the value of pixel (x_0,y_0) in the amplitude image A exceeds a valid range $[A_{\min},A_{\max}]$, which means that the reflectivity of the pixel is awfully weak or very strong in an abnormal setting.

5.2.2 Boundary reliability weight

Since the stereo matching typically experiences serious problems in the vicinity of 3D depth boundaries, a boundary reliability weight of ToF depth maps should be designed to indicate whether the estimated depth is from the boundaries. The motivation of the boundary reliability weight is to reduce the reliability of the ToF depth of the pixels on the boundaries. Given each pixel (x_0,y_0) in the ToF depth map D, the depth value of (x_0,y_0) in D is denoted as $D(x_0,y_0)$. The boundary reliability weight $W_{\mathrm{D}}(x_0,y_0)$ of (x_0,y_0) in D is defined as follows:

$$W_{\mathrm{D}}(x_0,y_0)=\mathrm{e}^{-S_{\mathrm{D}}(x_0,y_0)/\varphi}, \tag{3}$$

$$S_{\mathrm{D}}(x_0,y_0)=\sum_{(x,y)\in N(x_0,y_0)}|D(x,y)-D(x_0,y_0)|, \tag{4}$$

where $N(x_0,y_0)$ is a chosen neighboring window for boundary detection, $S_{\mathrm{D}}(x_0,y_0)$ is the average SAD of the depth difference measurement in the neighborhood $N(x_0,y_0)$, and φ is a certainty threshold value of the SAD of depth difference.

5.2.3 ToF energy cost function

The methods used in Diebel and Thrun (2005) and Gandhiy *et al.* (2012) sometimes blur the correct depth estimated by the ToF camera due to the global optimization algorithms used in these methods. To prevent such inaccuracy, since the ToF depth map can be converted into a parallax map, we define a parallax penalty cost to evaluate whether the parallax between one pixel and its correspondence pixel is suitable or not. The parallax penalty cost is defined as $\mathrm{e}^{|(x-x_0)-P_{\mathrm{D}}(x_0,y_0)|/\theta}$, where $P_{\mathrm{D}}(x_0,y_0)$ is the guiding parallax of ToF pixel (x_0,y_0) calculated with the intrinsic parameters of the RGB camera, and θ is a threshold. With this penalty cost, the chosen matching parallax should not be far from the reliable ToF depth parallax. At last, the ToF energy cost function can be defined as follows:

$$\begin{aligned}&C_{\mathrm{D}}(x,y_0)\\ =&W_{\mathrm{A}}(x_0,y_0)\cdot W_{\mathrm{D}}(x_0,y_0)\cdot \mathrm{e}^{|(x-x_0)-P_{\mathrm{D}}(x_0,y_0)|/\theta}.\end{aligned} \tag{5}$$

5.3 *K*-nearest neighborhood (KNN) based stereo matching with the energy cost function

Now the local stereo matching is not only relied on the matching cost calculated from two HD RGB cameras, but also helped by the ToF cameras. The updated matching cost C for stereo matching is defined as follows:

$$C(x,y_0)=C_{\mathrm{o}}(x,y_0)+C_{\mathrm{D}}(x,y_0), \tag{6}$$

where C_{o} is the original matching cost calculated by the traditional stereo matching method and C_{D} is the ToF energy cost in Eq. (5). This updated matching cost C is used to perform the local stereo matching. During the local stereo matching process, to merge the local pixels's costs with similar color distributions for more accurate matching, we also performed a KNN distance searching algorithm (Chen *et al.*, 2012) to determine the suitable local window size for stereo matching. While choosing the suitable window size for stereo matching, we defined a distance form between two pixels (x_0,y_0) and (x,y) in a local area as

$$D_n=||\boldsymbol{T}(x_0,y_0)-\boldsymbol{T}(x,y)||, \tag{7}$$

where $\boldsymbol{T}$ is a vector of pixel (x,y) containing five parameters: R, G, B components and x, y coordinates. This form of distance evaluates the similarity between a local pixel (x,y) and the target stereo matching pixel (x_0,y_0). Only the pixels with a D_n lower than a chosen threshold D_{T} will be selected in the local stereo match window for the matching process. The value of threshold D_{T} will be discussed in Section 7.

The local window based matching cost will choose pixel $(x_{\min},y_0)$ on the left image which has the minimum matching cost to be the matched pixel, so the matched parallax between these two matched pixels is $P_{\mathrm{r}}(x_0,y_0)=|x_{\min}-x_0|$. To calculate the new depth map D_{M}, the parallax map is converted

into a depth map by the intrinsic parameters of RGB cameras. This can be done as follows:

$$D_{\mathrm{M}}(x_0, y_0) = \frac{fB}{P_{\mathrm{r}}(x_0, y_0)}, \tag{8}$$

where f refers to the focal length of RGB cameras and B refers to the baseline between two RGB cameras. With the help of the ToF energy cost function, those regions with textureless or repetitive patterns can gain a more accurate depth. The ToF guided stereo matching procedure is described in Algorithm 1.

Algorithm 1 ToF guided stereo matching

```
COSTFUNCTION(x0, y0, p)
for (x0, y0) ∈ Ir do
    for (x, y0) ∈ [(x0 − m, y0), (x0 + m, y0)] do
        Calculate C(x, y0)
        Cmin = C(x, y0) ← min
    end for
    d(x0, y0) ← p with min C
end for
```

6 Depth fusion

Although we can obtain a more accurate high resolution depth map of the scene by the modified passive stereo with the help of data from ToF cameras, the acquired depth map is still sub-optimal due to the innate limitation of stereo matching. For example, the pixels on the right color image probably do not have correspondence pixels on the left color image. The estimated depth information by ToF cameras is also unsatisfactory since some fine-grained structures which typically have sufficient saliency are invisible in the ToF depth maps due to the low resolution of ToF cameras.

Therefore, it is desirable to borrow the strengths from stereo matching and ToF cameras. That is to say, we attempt to fuse the depth maps by stereo matching and ToF cameras together for the generation of a high accuracy depth map. To fuse two kinds of depth maps, we needed to find an appropriate optimization weighting schema. Depending on the weighting schema, for each pixel, and in a certainty parallax search range, the parallax with the minimum weighting value shall be the best parallax candidate. It is attractive to leverage necessary cues during fusion. Therefore, this study designs the optimization weight from a warped high resolution ToF depth map, a guided stereo matching depth map, and a high resolution referenced color image. Accordingly, two confidence weighting terms and one consistent weight (e.g., the confidence weights for ToF depth and stereo matching depth, the consistency for the referential color images) are respectively proposed to compose the final optimization weight for fusion.

6.1 Confidence weights of ToF depth

The degree of the approximation of the estimated ToF depth to the true depth, is used to indicate the reliability of a searching parallax p under the help of an estimated ToF depth value. For each searching parallax p, a larger ToF depth confidence weight implies a more unreliable parallax. After we convert the ToF depth map D into a parallax map P, the confidence weight $\mathrm{IC_D}$ for pixel (x_0, y_0) and each searching parallax p is defined as

$$\mathrm{IC_D}(x_0, y_0, p) = \begin{cases} \min(W_{\mathrm{A}} \cdot W_{\mathrm{D}} \cdot (p - P(x_0, y_0))^2, \alpha), \\ \qquad \text{if } p_{\min} \le p \le p_{\max}, \\ \alpha, \qquad \text{otherwise}, \end{cases} \tag{9}$$

where $P(x_0, y_0)$ is the estimated ToF depth estimation converted into the parallax disparity form, $(p_{\min}, p_{\max})$ is a chosen searching range of parallax disparity, and α is a threshold value of this ToF confidence cost. A better α can avoid the cost going to the maximum when the searching parallax p is out of the correct parallax disparity range or the ToF estimation is unreliable. The cost function $\mathrm{IC_D}$ also uses weights W_{A} and W_{D} which have been defined in the previous section for the defined ToF confidence weight.

6.2 Confidence weights of the stereo matching depth

Similar to the confidence weights of the ToF depth, the confidence weights of stereo matching depth $\mathrm{IC_M}$ implies the degree of reliability of each searching parallax p with respect to the stereo matching depth map. A naive intuition of checking whether the stereo matching result of pixel $D_{\mathrm{M}}(x_0, y_0)$ is correct or not is that the RGB value of pixel (x_0, y_0) on the right color image I_{r} nearly equals the RGB value

of pixel $(x_0 - P_{\mathrm{r}}(x_0, y_0), y_0)$ on the left color image. According to this intuition, the confidence weight of the stereo matching depth $\mathrm{IC_M}$ is defined as follows:

$$\mathrm{IC_M}(x_0, y_0, p) = \begin{cases} \min(I_{\mathrm{r}}(x_0, y_0) - I_{\mathrm{l}}(x_0 - p, y_0), \beta), \\ \qquad \text{if } p_{\min} \le p \le p_{\max}, \\ \beta, \qquad \text{otherwise}, \end{cases} \tag{10}$$

where $I_{\mathrm{r}}(x_0, y_0)$ and $I_{\mathrm{l}}(x_0 - p, y_0)$ refer to the RGB values of the pixels in the paired color images I_{r} and I_{l}, respectively. The SAD of $I_{\mathrm{r}}(x_0, y_0)$ and $I_{\mathrm{l}}(x_0 - p, y_0)$ in terms of R, G, and B color spaces is also calculated, respectively. β is a predefined threshold value (we will discuss β in the experiments), which is used to remove the adverse effects of large errors in stereo matching.

6.3 Normalization of confidence weights

We normalize the confidence weights of ToF depth $\mathrm{IC_D}$ and stereo matching depth $\mathrm{IC_M}$ into $[0, 1]$, referring to the normalized counterparts as C_{D} and C_{M}. They are defined as follows:

$$\begin{aligned} \overline{\mathrm{IC_D}(x_0, y_0)} &= \max_p(\mathrm{IC_D}(x_0, y_0, p)), \\ \underline{\mathrm{IC_D}(x_0, y_0)} &= \min_p(\mathrm{IC_D}(x_0, y_0, p)), \end{aligned} \tag{11}$$

$$C_{\mathrm{D}}(x_0, y_0, p) = \begin{cases} \frac{\mathrm{IC_D}(x_0, y_0, p) - \underline{\mathrm{IC_D}(x_0, y_0)}}{\overline{\mathrm{IC_D}(x_0, y_0)} - \underline{\mathrm{IC_D}(x_0, y_0)}}, \\ \quad \text{if } \overline{\mathrm{IC_D}(x_0, y_0)} - \underline{\mathrm{IC_D}(x_0, y_0)} \ge \Delta_{\mathrm{D_{noise}}}, \\ 1, \qquad \text{otherwise}, \end{cases} \tag{12}$$

$$\begin{aligned} \overline{\mathrm{IC_M}(x_0, y_0)} &= \max_p(\mathrm{IC_M}(x_0, y_0, p)), \\ \underline{\mathrm{IC_M}(x_0, y_0)} &= \min_p(\mathrm{IC_M}(x_0, y_0, p)), \end{aligned} \tag{13}$$

$$C_{\mathrm{M}}(x_0, y_0, p) = \begin{cases} \frac{\mathrm{IC_M}(x_0, y_0, p) - \underline{\mathrm{IC_M}(x_0, y_0)}}{\overline{\mathrm{IC_M}(x_0, y_0)} - \underline{\mathrm{IC_M}(x_0, y_0)}}, \\ \quad \text{if } \overline{\mathrm{IC_M}(x_0, y_0)} - \underline{\mathrm{IC_M}(x_0, y_0)} \ge \Delta_{\mathrm{M_{noise}}}, \\ 1, \qquad \text{otherwise}, \end{cases} \tag{14}$$

where $\Delta_{\mathrm{D_{noise}}}$ and $\Delta_{\mathrm{M_{noise}}}$ are the thresholds chosen to guarantee the least difference between the minimum and maximum costs. If the difference between the minimum and maximum costs is less than the threshold, the estimated depth at pixel (x_0, y_0) would be unreliable. The normalized C_{D} and C_{M} can effectively increase the sensitivity to noise in homogeneous regions of the images.

6.4 Consistency weight for referential images

Since each referential image I_{r} can provide subtle visual characteristics for its captured scene, the proposed fusion algorithm also takes the referential color image I_{r} into consideration. We create a color consistency weight for each pixel using a chosen local window. A large consistency weight value implies that the pixel is at the edge in the scene. This weight can recover some invisible fine-grained structures in the depth map which are visible in the color images. This weight contains two parts, the effectiveness weight $W_{I_{\mathrm{D}}}$ and the color difference weight $W_{I_{\mathrm{C}}}$. They are defined based on two observations: (1) It can be expected that if the spatial distance of the neighboring pixels increases, the effectiveness weight $W_{I_{\mathrm{D}}}$ will be reduced; (2) It can be assumed that depth discontinuities and color discontinuities should coincide. That is to say, neighboring pixels with different colors are expected to belong to different depth levels, so weight $W_{I_{\mathrm{C}}}$ is designed under this assumption. $W_{I_{\mathrm{D}}}$ and $W_{I_{\mathrm{C}}}$ are combined into a total weight W_I. These weights are defined in a value range of $[0, 1]$:

$$\begin{aligned} W_{I_{\mathrm{D}}}(x, y, x_0, y_0) &= \mathrm{e}^{-\|(x,y)-(x_0,y_0)\|_2/\phi_{\mathrm{D}}}, \\ W_{I_{\mathrm{C}}}(x, y, x_0, y_0) &= \mathrm{e}^{-\|I_{\mathrm{r}}(x,y)-I_{\mathrm{r}}(x_0,y_0)\|_2/\phi_{\mathrm{C}}}, \\ W_I(x, y, x_0, y_0) &= W_{I_{\mathrm{D}}}(x, y, x_0, y_0) \cdot W_{I_{\mathrm{C}}}(x, y, x_0, y_0), \\ & \qquad (x, y) \in N(x_0, y_0), \end{aligned} \tag{15}$$

where $N(x_0, y_0)$ is the chosen neighborhood of pixel (x_0, y_0), and ϕ_{D} and ϕ_{C} keep the effectiveness weight $W_{I_{\mathrm{D}}}$ and the color difference weight $W_{I_{\mathrm{C}}}$ for the referential image in a reasonable range.

6.5 Optimization weight

Now the optimization weight C_{F} for each pixel and each parallax is defined as follows:

$$\begin{aligned} & C_{\mathrm{F}}(x_0, y_0, p) \\ &= \sum_{(x,y)\in N(x_0,y_0)} [W_I(x, y, x_0, y_0) \\ &\quad \cdot (C_{\mathrm{D}}(x, y, p) + C_{\mathrm{M}}(x, y, p))] \\ &\quad / \sum_{(x,y)\in N(x_0,y_0)} W_I(x, y, x_0, y_0). \end{aligned} \tag{16}$$

We look for the minimum value of C_{F} in a chosen searching parallax area $(p_{\min}, p_{\max})$. The parallax p_{F} which has the minimum value of C_{F} will be the correct parallax of pixel (x_0, y_0). After the parallax map is acquired for all of pixels, the obtained parallax map is converted into a depth map using the intrinsic parameters of RGB cameras. By this way, we can obtain the final high resolution depth map with high accuracy. The fusion procedure is described in Algorithm 2.

Algorithm 2 Depth fusion

FUSIONCOST(x_0, y_0, p)
for $(x_0, y_0) \in I_{\mathrm{r}}$ **do**
  for $p \in (p_{\min}, p_{\max})$ **do**
    Calculate $C_{\mathrm{F}}(x_0, y_0, p)$
    $C_{\mathrm{F}_{\min}} = C(x_0, y_0, p) \leftarrow \min$
    Chosen $p \leftarrow C_{\mathrm{F}_{\min}}(x_0, y_0, p)$
  end for
  $d(x_0, y_0) \leftarrow p$ with min C_{F}
end for

7 Experiments

In this section we evaluate the proposed approach on both benchmark image pairs and the images captured in the lab with high complexity scenes. We focus on the evaluation of the accuracy of obtained depth maps.

7.1 Experimental setup

For the benchmark test image pairs, since the ground-truth depth maps of the corresponding images are available (Scharstein and Szeliski, 2002b), we use the ground-truth depth maps to synthesize its ToF depth images. The ground-truth depth maps have high resolutions (i.e., 370×370), the same as the pairs of color images. We perform downsampling for ground-truth depth maps and the downsampled depth maps with a resolution of 200×200 are taken as the depth map gained by ToF cameras. Then we randomly add noise to some regions of the downsampled depth maps. Figs. 3b and 3c show one ground-truth depth map and its corresponding synthesized depth image. Then the proposed approach uses the pairs of color images and the downsampled depth map to obtain a high resolution depth map. To demonstrate the advantage of the proposed method, the results carried out by other stereo matching methods and depth fusion methods (Zhu *et al.*, 2008; Zhang *et al.*, 2010) are compared in terms of the root mean squared error (RMSE) of the ground truth, where RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum\limits_{(x,y)\in D} (D(x,y) - D_{\mathrm{G}}(x,y))^2}{N_{\mathrm{D}}}}, \quad (17)$$

where D is the fusion depth map, D_{G} is the ground-truth map, and N_{D} is the total pixel number of the depth map.

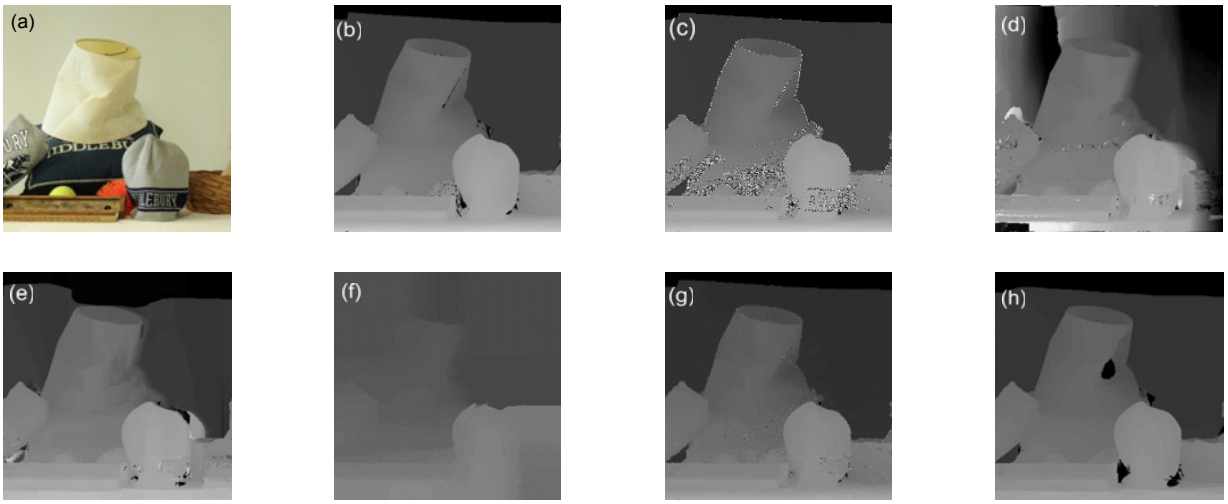

Fig. 3 Results and comparisons of experimental benchmark data: (a) right view image; (b) ground-truth depth map; (c) synthesized ToF depth map; (d) the result by traditional stereo matching (Zhang *et al.*, 2010); (e) the result by our ToF guided stereo matching; (f) the result by MRF fusion (Zhu *et al.*, 2008); (g) the result by our fusion with no ToF depth confidence weight; (h) the result by our proposed fusion approach

The proposed fusion approach is also tested on two datasets of images captured in our lab with high complexity scenes. Each dataset consists of two color view images with a high resolution of 540×540 pixels, and one low resolution ToF depth map (200×200 pixels). Figs. 4a, 4b, 5a, and 5b show the images captured by the cameras in our lab. The method described in Zhang (1999) was used to find the intrinsic and extrinsic parameters of cameras, thereafter the stereo depth map was estimated for the referenced color view image, the ToF depth map and ToF amplitude map are warped into the same resolution of this color view image according to the description in Section 3. Then the ToF guided stereo matching and depth fusion are implemented to obtain high accuracy depth maps respectively. To demonstrate the advantage of the appropriate fusing of the necessary cues, we also report the results when the ToF depth confidence weight is neglected in the proposed fusion approach.

To find a close approximation to the depth values with the minimal function costs, the function was densely sampled using an equally spaced hypothesis in the search range from $Z_{\min} = 1$ m to $Z_{\max} = 7.5$ m

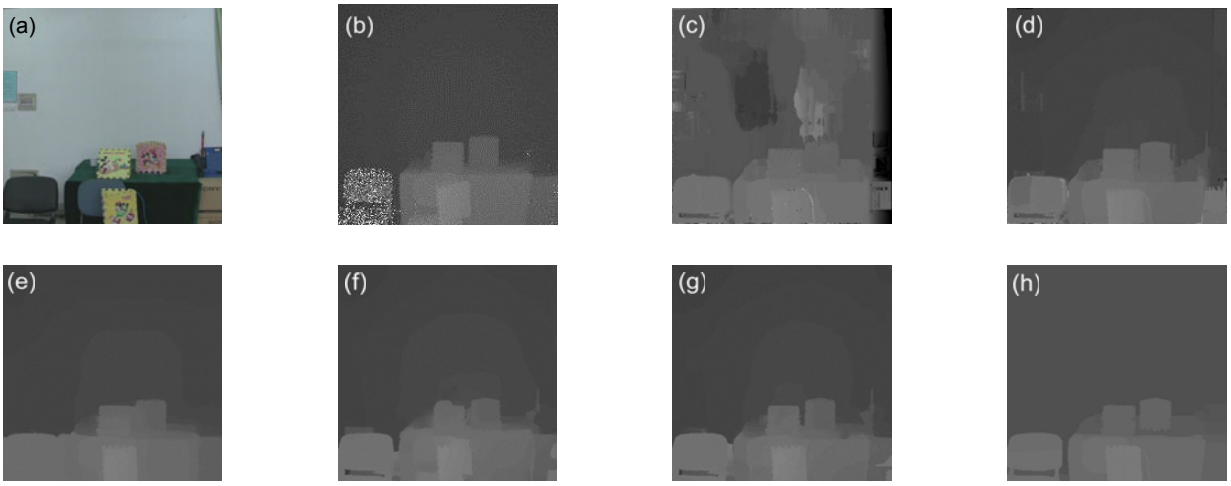

Fig. 4 Results and comparisons of captured data in our lab (data1): (a) right view image; (b) ToF depth map; (c) result by traditional stereo matching (Zhang *et al.*, 2010); (d) the result by our ToF guided stereo matching; (e) the result by MRF fusion (Zhu *et al.*, 2008); (f) the result by our fusion with no ToF depth confidence weight; (g) the result by our proposed fusion approach; (h) the manually annotated ground truth

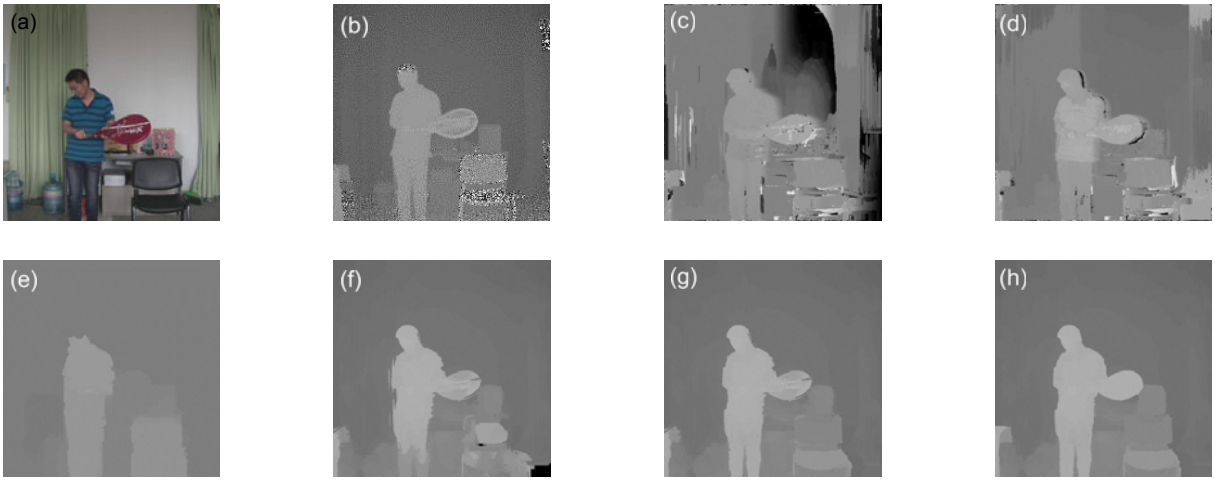

Fig. 5 Results and comparisons of captured data in our lab (data2): (a) right view image; (b) ToF depth map; (c) result by traditional stereo matching (Zhang *et al.*, 2010); (d) the result by our ToF guided stereo matching; (e) the result by MRF fusion (Zhu *et al.*, 2008); (f) the result by our fusion with no ToF depth confidence weight; (g) the result by our proposed fusion approach; (h) the manually annotated ground truth

of the ToF cameras. Then we calculated the search area of the parallax disparity $[p_{\min}, p_{\max}]$ with the intrinsic and extrinsic parameters of the RGB cameras. After selecting the parallax sample p with the lowest costs of function C_{D}, the same sampling was used to determine approximations to the minimal and maximal matching costs $\overline{\mathrm{IC}_{\mathrm{M}}(x_0, y_0)}$ and $\underline{\mathrm{IC}_{\mathrm{M}}(x_0, y_0)}$. The final depth approximation was found by computing the minimum of a quadratic polynomial C_{F} which is defined in the previous section.

7.2 Parameters

The parameters we defined in previous sections are important for the proposed approach. We tune these parameters experientially through a series of experimental datasets. We use a set of indoor image pairs to tune these parameters. Both the benchmark and our lab's image pairs are used for the parameter tuning process. We classify these image pairs into three categories: (1) the scenes with only long range objects (about 3–5 m from the camera), (2) the scenes with only short range objects (about 1–3 m from the camera), (3) the scenes with complex objects (both short and long range). We test and tune the parameters with these three categories of datasets. The performance of the parameters defined above is different from category to category. δ and γ determine the sensitivity of the confidence of the ToF depth map. The chosen value in Table 7.2 of δ and γ can choose the suitable reliable area of the ToF depth map. The chosen φ and θ can make C_{o} and C_{D} be in the same value range. α and β are the thresholds for the confidence weights of the ToF depth and stereo match depth, respectively. A larger or a smaller threshold will cause either a boundary destruction or reliable depth loss. ϕ_{D} and ϕ_{C} keep the consistency weight for the referential image in a reasonable range. And the noise thresholds D_{T}, $\Delta_{\mathrm{D}_{\mathrm{noise}}}$, and $\Delta_{\mathrm{M}_{\mathrm{noise}}}$ are chosen based on the average noise level of the ToF depth image. The values with the best performance of all three categories of datasets are chosen in the proposed approach. The parameters we used in the experiments are shown in Table 1.

Table 1 Parameters used in experiments

Parameter	Value	Parameter	Value
δ	10	γ	12
φ	30	θ	3
α	20	β	60
ϕ_{D}	50	ϕ_{C}	30
D_{T}	20	$\Delta_{\mathrm{D}_{\mathrm{noise}}}, \Delta_{\mathrm{M}_{\mathrm{noise}}}$	10

7.3 Comparisons and discussions

Fig. 3 shows the results of the benchmark image dataset. The depth maps obtained in Figs. 3d and 3f are calculated using the methods mentioned in Zhu *et al.* (2008) and Zhang *et al.* (2010), respectively. Fig. 3 shows the depth maps obtained by the proposed ToF guided stereo matching and fusion approach. Figs. 3–5 also demonstrate the obtained depth maps for the captured scenes in our lab by the proposed approach and other counterpart approaches. It should be pointed out that the ground-truth depth map of our captured scene is manually annotated by ourselves.

Due to the introduction of the high resolution referential color image data into the depth fusion in the proposed approach, it was shown that the proposed weights design can reliably preserve the subtleness of high resolution structures. Figs. 3g, 4f, and 5f show the results obtained by the proposed fusion approach when the ToF depth confidence weight is disregarded. It can be seen that the obtained results have some noisy regions due to the lack of the ToF depth confidence weight. As a result, an appropriate integration of necessary cues is very important for high accuracy depth maps.

Table 2 provides a summary of RMSE performances for the compared algorithms on benchmark data and our captured data in the lab. The benchmark data has a range of only 0.5–1.5 m and its maximum parallax is 20 pixels. The data from our captured scene has a range of 1–5 m and its maximum parallax is about 80 pixels.

From the results in Table 2, we can make the following observations:

The proposed fusion approach achieves the best performance of the acquired depth maps in terms of RMSE for benchmark data and our captured scene. As a result, it is beneficial to integrate the ToF cameras and traditional stereo matching in the process of obtaining high accuracy depth maps.

The appropriate introduction of the ToF depth map is very important in traditional stereo matching, although the MRF fusion method (Zhu *et al.*, 2008) does introduce the BP aggregation method into the

Table 2 Summary of performance for the compared algorithms on benchmark data and captured data in our lab in terms of RMSE

Dataset	Method	Resolution	RMSE (m)	Time (ms)
Benchmark data	Original depth map (ground truth)	370 × 370	0.029	null
Benchmark data	Tractional stereo matching (Zhang *et al.*, 2010)	370 × 370	0.067	75
Benchmark data	ToF guided stereo matching	370 × 370	0.054	77
Benchmark data	MRF fusion (Zhu *et al.*, 2008)	370 × 370	0.094	89
Benchmark data	Our proposed fusion without ToF depth confidence weight	370 × 370	0.045	**62**
Benchmark data	Our proposed fusion	370 × 370	**0.035**	69
Our lab data1	Original depth map	540 × 540	0.167	null
Our lab data1	Tractional stereo matching (Zhang *et al.*, 2010)	540 × 540	0.122	112
Our lab data1	ToF guided stereo matching	540 × 540	0.105	120
Our lab data1	MRF fusion (Zhu *et al.*, 2008)	540 × 540	0.103	185
Our lab data1	Our proposed fusion without ToF depth confidence weight	540 × 540	0.098	**98**
Our lab data1	Our proposed fusion	540 × 540	**0.094**	105
Our lab data2	Original depth map	540 × 540	0.188	null
Our lab data2	Tractional stereo matching (Zhang *et al.*, 2010)	540 × 540	0.131	114
Our lab data2	ToF guided stereo matching	540 × 540	0.112	121
Our lab data2	MRF fusion (Zhu *et al.*, 2008)	540 × 540	0.108	183
Our lab data2	Our proposed fusion without ToF depth confidence weight	540 × 540	0.100	**99**
Our lab data2	Our proposed fusion	540 × 540	**0.093**	104

The highest performance is highlighted in each case. The running time performance is achieved on a standard desktop with 3.2 GHz CPU and 8 GB RAM

fusion of the ToF depth and stereo matching depth. However, in some cases such methods do not suffice to prevent a bias towards a wrong depth hypothesis. The situation would get worse if the wrong hypothesis is additionally supported by the depth estimation of the ToF cameras. In our proposed approach, two mechanisms were considered to prevent such a wrong hypothesis. First, both the confidence weights of the ToF depth and stereo matching depth are normalized into $[0, 1]$. By this way, the consistency of the pixel based referential color image can compete with the depth weights. Second, we integrate the consistency distance weight of the referential color image into adaptive optimization (Section 6), so that discontinuities can be handled properly even if the required color contrast is very low.

The ToF guided stereo matching also achieves a better performance than traditional stereo matching in terms of RMSE for benchmark data and our captured scene. We can observe from Figs. 3d, 3e, 4c, 4d, 5c, and 5d that the depths of textureless and repetitive regions in the images have been observably improved.

8 Conclusions

In this paper we propose a robust approach to obtaining high-accuracy depth maps by the integration of passive stereo and ToF cameras. The proposed approach can exploit two color images and a ToF depth image to obtain a high resolution depth map. The main contributions of this study are as follows: (1) An energy cost function is devised to utilize data from ToF cameras to boost the stereo matching of passive stereo; (2) A fusion method is used to combine the depth information from both ToF cameras and passive stereo to obtain high accuracy depth maps. The experiments show that the proposed approach achieves improved results with high accuracy and robustness.

References

Attamimi, M., Mizutani, A., Nakamura, T., *et al.*, 2010. Real-time 3D visual sensor for robust object recognition. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, p.4560-4565. [doi:10.1109/IROS.2010.5650455]

Buehler, C., Bosse, M., McMillan, L., *et al.*, 2001. Unstructured lumigraph rendering. Proc. 28th Annual Conf.

on Computer Graphics and Interactive Techniques, p.425-432. [doi:10.1145/383259.383309]

Canesta, 2006. Canestavision Electronic Perception Development Kit. Available from http://www.canesta.com/

Chen, Q., Li, D., Tang, C., 2012. KNN matting. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.869-876. [doi:10.1109/CVPR.2012.6247760]

De-Maeztu, L., Mattoccia, S., Villanueva, A., *et al.*, 2011. Linear stereo matching. Proc. IEEE Int. Conf. on Computer Vision, p.1708-1715. [doi:10.1109/ICCV.2011.6126434]

Diebel, J., Thrun, S., 2005. An application of Markov random fields to range sensing. *Adv. Neur. Inform. Process. Syst.*, **18**:291-298.

Gandhiy, V., Cech, J., Horaud, R., 2012. High-resolution depth maps based on TOF-stereo fusion. IEEE Int. Conf. on Robotics and Automation, p.4742-4749. [doi:10.1109/ICRA.2012.6224771]

Gudmundsson, S.A., Aanaes, H., Larsen, R., 2008. Fusion of stereo vision and time-of-flight imaging for improved 3D estimation. *Int. J. Intell. Syst. Technol. Appl.*, **5**(3-4):425-433. [doi:10.1504/IJISTA.2008.021305]

Kanade, T., Okutomi, M., 1994. A stereo matching algorithm with an adaptive window: theory and experiment. *IEEE Trans. Patt. Anal. Mach. Intell.*, **16**(9):920-932. [doi:10.1109/34.310690]

Lee, C., Song, H., Choi, B., *et al.*, 2011. 3D scene capturing using stereoscopic cameras and a time-of-flight camera. *IEEE Trans. Consum. Electron.*, **57**(3):1370-1376. [doi:10.1109/TCE.2011.6018896]

Lindner, M., Kolb, A., Hartmann, K., 2007. Data-fusion of PMD-based distance-information and high-resolution RGB-images. Proc. Int. Symp. on Signals, Circuits and Systems, p.1-4. [doi:10.1109/ISSCS.2007.4292666]

May, S., Werner, B., Surmann, H., *et al.*, 2006. 3D time-of-flight cameras for mobile robotics. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, p.790-795. [doi:10.1109/IROS.2006.281670]

Opencv, 2012. Open Source Computer Vision Library (opencv). Available from www.intel.com/technology/computing/opencv/

PMD, 2009. Camcube Series. Available from http://www.pmdtec.com/

PMD, 2010. Camcube 3.0 Products. Available from http://www.pmdtec.com/products-services/pmdvisionr-cameras/pmdvisionr-camcube-30/

Ringbeck, T., Hagebeuker, B., 2007. A 3D time of flight camera for object detection. Proc. 8th Conf. on Optical 3-D Measurement Techniques.

Scharstein, D., Szeliski, R., 2002a. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.*, **47**(1-3):7-42. [doi:10.1023/A:1014573219977]

Scharstein, D., Szeliski, R., 2002b. Middlebury Stereo Evaluation - version 2. Available from http://vision.middlebury.edu/stereo/eval

Wang, L., Lou, L., Yang, C., *et al.*, 2013. Portrait drawing from corresponding range and intensity images. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **14**(7):530-541. [doi:10.1631/jzus.CIDE1306]

Xu, Z., Schwarte, R., Heinol, H., *et al.*, 1998. Smart pixel-photonic mixer device (PMD). Proc. 5th Int. Conf. on Mechatronics and Machine Vision in Practice, p.259-264.

Yang, Q., 2012. A non-local cost aggregation method for stereo matching. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1402-1409. [doi:10.1109/CVPR.2012.6247827]

Yao, L., Li, D., Zhang, J., *et al.*, 2012. Accurate real-time stereo correspondence using intra- and inter-scanline optimization. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **13**(6):472-482. [doi:10.1631/jzus.C1100311]

Z-cam, 2004. 3DV Systems. Available from http://www.3dvsystems.com

Zhang, J., Li, D., Zhang, M., 2010. Fast stereo matching algorithm based on adaptive window. Proc. Int. Conf. on Audio Language and Image Processing, p.138-142. [doi:10.1109/ICALIP.2010.5684994]

Zhang, Z., 1999. Flexible camera calibration by viewing a plane from unknown orientations. Proc. 7th IEEE Int. Conf. on Computer Vision, p.666-673. [doi:10.1109/ICCV.1999.791289]

Zhu, J., Wang, L., Yang, R., *et al.*, 2008. Fusion of time-of-flight depth and stereo for high accuracy depth maps. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-8. [doi:10.1109/CVPR.2008.4587761]