Reliability Fusion of Time-of-Flight Depth and Stereo Geometry for High Quality Depth Maps

Jiejie Zhu, *Member*, *IEEE*, Liang Wang, *Student Member*, *IEEE*, Ruigang Yang, *Member*, *IEEE*, James E. Davis, *Member*, *IEEE*, and Zhigeng Pan, *Member*, *IEEE*

Abstract—Time-of-flight range sensors have error characteristics, which are complementary to passive stereo. They provide real-time depth estimates in conditions where passive stereo does not work well, such as on white walls. In contrast, these sensors are noisy and often perform poorly on the textured scenes where stereo excels. We explore their complementary characteristics and introduce a method for combining the results from both methods that achieve better accuracy than either alone. In our fusion framework, the depth probability distribution functions from each of these sensor modalities are formulated and optimized. Robust and adaptive fusion is built on a pixel-wise reliability weighting function calculated for each method. In addition, since time-of-flight devices have primarily been used as individual sensors, they are typically poorly calibrated. We introduce a method that substantially improves upon the manufacturer's calibration. We demonstrate that our proposed techniques lead to improved accuracy and robustness on an extensive set of experimental results.

Index Terms—Time-of-Flight sensor, multisensor fusion, global optimization, stereo vision.

1 Introduction

Depth sensing is one of the fundamental challenges of computer vision. Applications include robotic navigation, object reconstruction, and human computer interaction. A range sensor that is robust, accurate, and operates in real time would be the enabling component in these applications. Unfortunately, no existing range sensing method is perfect on its own. For example, laser scanners are too slow for real time usage, passive stereo can easily fail on textureless scenes, photometric stereo is prone to low-frequency distortion, and only-recently-available Time-of-Flight (ToF) sensors are low in resolution, noisy, and poorly calibrated.

The ToF sensor provides real-time independent range estimates at each pixel, and has only recently started to become available from companies such as Canesta [1], SwissRanger [2], 3DV [3], and PMD [4] at commodity prices. Due to their recent introduction, most applications use the sensors individually and rely on the manufacturer's calibration. Despite their promise, relatively little literature explores the ways in which the quality of these sensors might be improved.

- *J. Zhu and Z. Pan are with the Digital Media and Human Computer Intersection Research Center (DMHCIRC), Hangzhou Normal University, Hangzhou, China 310036. E-mail: jjzhu@cs.ucf.edu, zgpan@cad.zju.edu.cn.*
- *L. Wang is with the Microsoft Applied Science Group, Microsoft Corp., One Microsoft Way, Redmond, WA 98121. E-mail: lwan@cs.uky.edu.*
- *R. Yang is with the Computer Science Department, University of Kentucky, 1 Quality Street, Suite 800, Lexington, KY 40507-1464. E-mail: ryang@cs.uky.edu.*
- *J.E. Davis is with the Center for Entrepreneurship at UCSC, 1156 High St #SOE3 UC Santa Cruz, CA 95064. E-mail: davis@cs.ucsc.edu.*

Manuscript received 26 Aug. 2008; revised 5 July 2010; accepted 19 July 2010; published online 31 Aug. 2010.
Recommended for acceptance by J. Oliensis.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-08-0568.
Digital Object Identifier no. 10.1109/TPAMI.2010.172.

This paper seeks to improve the range estimates provided by the ToF sensor by combining it with both the ToF and relatively more sophisticated algorithms common to passive stereo vision. The ToF sensor is characterized by independent pixel range estimates, each of which has a relatively high noise that is difficult to model around the true depth because many factors may be present. Optical imperfection and scene reflectance are two main noise resources. Many of the industry calibration works have reported improvements on removing optical issues by testing error models. Unlike these approaches, we utilize structured light method to compute the ground truth depth and compare that with the depth reported by the ToF sensor. This simple, yet effective, nonparametric method allows us to measure the optical errors as well as depth bias, without the need for error models.

Our success in fusing the ToF sensor with passive stereo vision is based on their complementary nature. Rich texture causes difficulties for the ToF sensors because these sensors frequently have biases as a function of object albedo. Conversely, passive stereo excels on such regions as a unique local minima appears in the cost volume. However, passive stereo performs poorly on textureless regions, repeated patterns, and occluded areas, which will cause multiple local minima. In such regions, the ToF sensors can exceed passive stereo by calculating the time delay, while the emitted light is reflected back from objects. We explore this complementary nature to fuse the probability distribution functions of the depth estimates from each of these sensor modalities by using a Markov Random Field (MRF), which can produce a combined sensor with superior characteristics.

In this MRF model, the ToF sensor can be regarded as providing local data, while the passive stereo methods are quite sophisticated and have been carefully categorized according to the effects of changing the local matching function, aggregation function, and global regularization

[5]. We apply belief propagation (BP) to perform global optimization on the combined sensor and show that global optimization improves the ToF sensor's accuracy, just as it can improve passive stereo vision.

Our motivation to combine multiple sensors requires that they be calibrated in a common coordinate frame. Unfortunately, the ToF sensor is usually designed only to provide relative depth, as opposed to measurements in a calibrated euclidean frame. The main contributions of this paper are two-fold: 1) a method to calibrate the ToF sensor and passive stereo into a common euclidean coordinate system, and 2) a method for using data from both the ToF sensor and passive stereo to produce enhanced depth estimates by global regularization. Results show that the combined sensor can reduce the depth error from 1.8 percent to 0.6 percent in a 1.5 meter distance range. Although this distance range is limited, we envision the proposed calibration and fusion approach can also be employed for long distance range.

2 Related Works

There are many ways to obtain scene depth. In general, they can be categorized into two major classes: passive methods and active methods. Among the plethora of passive methods, stereovision [6] is probably the most well known, least expensive, and most widely used. It is beyond the scope of this paper to provide a brief review of existing stereo methods. Interested readers are referred to an excellent review by Scharstein and Szeliski [5]. Despite significant progress made during the last few years, the fundamental problems in stereo, such as occlusion, textureless, and repetitive patterns, remain unsolved.

The ToF sensors use an active technique to obtain near real-time scene depth. They are able to produce a full depth frame simultaneously, thereby allowing applications to dynamic scenes. There are mainly two types of ToF sensors. One utilizes modulated and incoherent light, which is based on phase shift that can be designed using standard CMOS or CCD technology [7], [8], [9]. The other is based on an optical shutter technology [10], [11]. The one we use belongs to the first category of ToF sensors. Differently from 3D scanning and structured coded light approach [12], [13], the ToF sensor can return a full frame depth measurement in real time instead of a point or a scan line.

The basic principle of phase shift is based on measuring the phase delay of the reflected light. The ToF sensors can return two types of data given the emitted and the returned signal: a depth map from phase shift and an intensity image from the amplitude. Note that some of the ToF sensors can also return a 3D point cloud of the scene, and the depth map is treated as the Z value.

Due to the complexity of the optical system and the real scene, the quality of the depth returned from the ToF sensors is subject to a number of noise factors. In the optical system, the main noise source is the photon shot noise that is theoretically Poisson distributed, and inhomogeneities in the near-infrared light filed of the LED array are also reported to disturb the depth measurement [14]. Internal and external temperatures are also observed to influence the depth measurement [15]. Additionally, other noise factors, such as multiple reflection, light scattering, glossy reflection, ambient light, and color difference have side effects to depth quality.

Previous works of calibration methods [16], [25], [18], as well as in-pixel background light suppression [8] and denoising by nonlocal median filter [19], have proven to be useful to reduce such noises. However, these methods have primarily looked at the specified error model of the ToF sensors that are hard to correctly model. Some of the methods use the depth reported from the ToF sensor alone to calibrate itself, which is problematic. Unlike these approaches, our approach is a nonparametric method. It requires no explicit error models which provide a general calibration method that compensates the depth bias for real scenes without assuming specifics of the ToF sensor. Compared with several methods to obtain ground truth depth, such as using a total station to measure coordinates of an object [15] or resorting to the trackline [25] and a robotic arm to determine ground truth depth [18], we obtain ground truth depth from structured light method, which is in high accuracy.

The depth maps returned from the ToF sensor are commonly in low resolution. This makes them less appealing for most vision algorithms. Several super-resolution methods [20], [21], [22] have been introduced to enhance its resolution. They are mainly based on the fact that discontinuities in range and coloring intend to coalign [23]. Our approach can do upsampling coherently, as we calibrate the ToF sensor and two stereo cameras into a common coordinate system. Using any one of the cameras as a reference view, we can obtain high resolution depth maps.

This paper does not intend to resolve all issues of the ToF sensor, such as the capture frequency. Readers may refer to [24] for details on how to increase it by a CMOS-based technique.

There are already several approaches to merge the ToF sensor with images captured from monochromatic or stereo cameras. In [25], a projective texture technique is used to align depth from the ToF sensor to a pixel on an RGB camera. In [26], [27], depth accuracy is improved by merging regional selected depth from stereo matching and depth from the ToF sensor. The comparison of depth accuracy between the ToF sensor and stereo rig is reported by Beder et al. [28]. The most similar setup to ours is [29], which aims to improve the depth by finding a dense correspondence between the stereo rig and the ToF sensor.

Another contribution of our work is that most of the previous methods are not able to give a metric report of the quantity of improvement compared to the state-of-art methods in stereo. In this paper, we obtain depth maps, both from state-of-art stereo methods and a structured light method. Thus, a quantitative evaluation on a number of real scenes is available by calculating the numerical difference between our approach and the ground truth.

3 Multi-Sensor Setup

In this paper, we set up a ToF sensor with a pair of cameras (as shown in Fig. 1). The stereo cameras have a baseline around 100 mm, and they are verged toward each other around 10 degrees from the parallel setup. Our setup is designed to provide coverage at about one meter range.

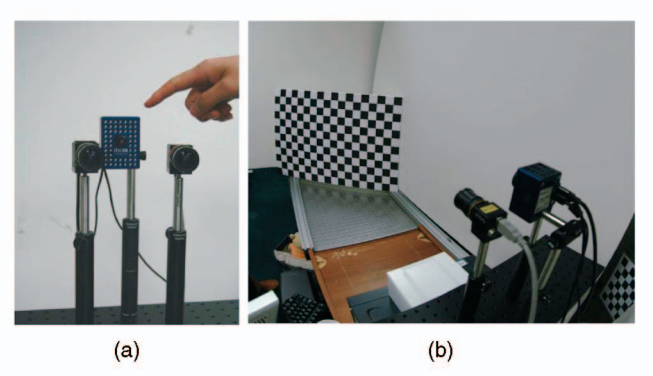


Fig. 1. (a) Multisensor setup with two CCD color cameras and one Swissranger SR3000. (b) Calibration setup. The rails on the table are used to move the pattern.

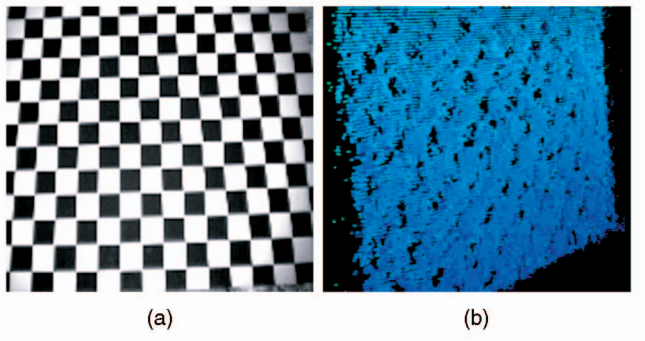


Fig. 2. Example of depth bias from a white/black planar object. (a) The intensity image returned from the ToF sensor. (b) The plane in 3D rendered using depth returned by the ToF sensor. Note the black holes are dark regions in (a).

Note that the 1 m range distance is not defined as the distance from the object to the ToF sensor, but as the distance inside the calibrated volume.

The ToF sensor we have is a SwissRanger SR3000 [2], which can continuously produce a depth map with a resolution of 176×144. Its operational range is up to 7.5 m with the modulation frequency set to 20MHZ. In addition, SR3000 will also produce an intensity image in the same resolution based on amplitude. Together with two color cameras, these three sensors can be calibrated into a common coordinate system using the traditional calibration method, which will be introduced in Section 4.3.

4 Multisensor Calibration

In this section, we introduce an empirical calibration method to improve the depth accuracy, using our setup.

Our approach generates per-pixel Look-Up-Tables (LUTs) to compensate the depth bias caused by various scene reflectance (photometric calibration) and system noise (geometric calibration). Given these LUTs, we first correct the depth returned from the ToF sensor and then refine it by merging with the depth from passive stereo.

Our calibration volume is within a 400 mm distance range, which is sampled with 16 distance steps, and the distance gap is around 25 mm. Experimentally, we notice that the depth bias caused by scene reflectance variations can be described using a per-pixel piece-wise linear function.

These linear functions are approximated using a black/white chessboard planar pattern. The linear function is estimated for each sampled distance. In the geometric calibration, we first calibrate the ToF sensor and stereo cameras into a common coordinate system. In this space, we can easily refine the depth from the ToF sensor, given the reference depth from the stereo. We now introduce these two processes, in detail, in the following sections.

4.1 Photometric Calibration

We observed that the depth returned from the ToF sensor is sensitive to different object reflectance. Fig. 2 shows an example of this problem, using a chessboard planar object.

Previous work on photometric calibration either uses median filter [30] or spline functions [25] to account for this bias. The depth of the centered pixel is estimated based on the assumption that the accuracy of depth correlates with intensities in a small neighborhood. In our approach, to compensate for depth bias from various scene reflectance, we estimate a piece-wise linear function via black and white planar calibration objects inside the calibrated volume.

To obtain the minimal and maximal scene reflectance from the ToF sensor, we placed these two planar objects at the largest distance (near 1.7 m) inside our calibrated volume, and the ToF sensor reports an average intensity value of 25 for the black planar object and 205 for the white planar object.

Then, we put the calibration objects at 16 sampled distances and plot the depth estimation between black/white planar object from each pixel. We notice that there is an almost constant shift between measurements at each sampled distance. Based on this observation, we build a per-pixel depth bias LUT according to the scene reflectance. Given the ToF sensor's measurement of certain scene points, its range bias is linearly interpolated based on its intensity difference to the black and white reference intensity values.

In Fig. 3a, we find that the piece-wise model is nearly a constant scalar, which fits well in our work to account for the depth bias from various scene reflectance. Our calibrated volume, although limited (from around 1.2 m to 1.6 m), has

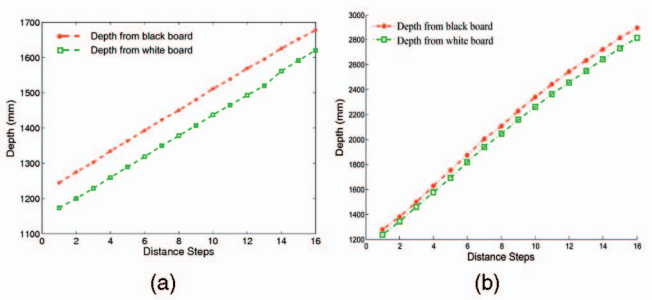


Fig. 3. Example of depth estimation at pixel (72, 75) from white/black planar objects in 16 sampled distances. The *Distance steps* on the horizontal axis denotes the number of sampled distance. The *Distance* on the vertical axis denotes the depth in the scene read out from the ToF sensor. At each sampled distance, the ToF sensor measured 10 times and the averaged depth estimation is used. The resolution of the ToF sensor is 176×144. In (a), the calibration range is around 1.0-1.7 m. In (b), the range is extended to around 3 m. (a) Limited calibration range. (b) Extended calibration range.

two benefits to our fusion approach: 1) We do not need to change the integration time. The integration time is similar to the shutter speed of a digital camera. The shorter the integration time, the fewer the accumulated photons, which leads to a larger depth bias. The difficulty in changing the integration time lies in the nonlinear depth bias, which is difficult to calibrate. 2) The depth bias inside each distance interval is very small. In photometric calibration, we divide the volume into 16 distance intervals, which is only 2.5 cm for each step. Our results show that the depth bias in each interval is almost the same as the previous interval. This demonstrates that a constant assumption of depth bias in each small interval is reasonable. Additionally, by successfully fusing information from passive stereo, the noise from the ToF sensor is suppressed and the depth bias is compensated for by utilizing neighboring information. Compared with other complex error models, such as [25], [18], our model is simple. And more importantly, it is nonparametric—we do not rely on specific error models.

We also test the depth difference, using black/white planar board for an extended range in Fig. 3b. The integration time controls the exposure time and can be varied from 200 μs to 51.2 ms in steps of 200 μs, whereas 0 corresponds to an integration time of 200 μs and 255 of 51.2 ms. We change the integration time three times: 40 for 1.2 m-1.6 m, 50 for 1.6 m-2.2 m, and 60 for 2.2 m-2.8 m. We find that although the depth difference is not a constant across the entire range, its change is smooth and slow. Therefore, a piece-wise linear-interpolation model is a good approximation.

4.2 Geometrical Calibration

For each depth map from the ToF sensor, we first perform photometric correction according to its intensity values, then apply the geometric correction introduced in this section.

The ToF sensor we used (SR3000) returns a 3D point cloud of the scene (its z value is the scene depth) and an intensity image. By assuming an *orthogonal projection* from the point cloud to the image, we can correspond each pixel coordinate to a 3D point. The *orthogonal projection* may not be exact because the camera lens performs a perspective projection. However, the ToF's camera system can be regarded as a weak perspective projection which can be approximated using an orthogonal projection plus a scale. In addition, the range variation of the calibration board compared to the range between the board and the camera is small. Therefore, the approximation of orthogonal projection can be used.

Differently from the ToF sensor, a disparity map is normally used to visualize the scene depth in the stereo literature. The disparity map is defined as the horizontal pixel displacements between the rectified left and right images. We use the left camera as the reference camera. In the fusion(Section 6), the cost of the stereo term is defined based on disparity, while the cost of the ToF term is based on the depth. We first convert the disparity in the stereo space to depth, and then compare it with the depth from the ToF sensor. Additional notations used in the geometric calibration are introduced in the following paragraph.

Fig. 4 shows the concept of our geometric calibration. X_{stereo} denotes triangulated 3D points from passive stereo. We denote the passive stereo's coordinate system as *stereo CS*. By treating the ToF sensor as a regular camera, we can calibrate it with stereo cameras. We denote its local coordinate system as *world CS*. We call the calibrated ToF sensor with the stereo *ToF camera*. The 3D points in this space are represented by X_{world}. We refer to the point cloud returned by the SR3000 as X_{ToF}, and they belong to the raw ToF sensor's coordinate system *ToF CS*.

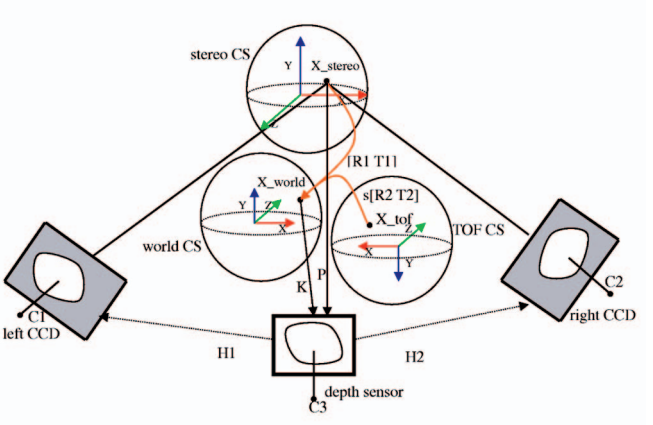


Fig. 4. Principle of geometric calibration. We calibrate the ToF sensor and stereo cameras into the *world CS* and compute the per-pixel depth bias by comparing depth from the ToF sensor and that from stereo cameras. In the *world CS*, P is the precalibrated projection matrix of the ToF sensor. K is the intrinsic matrix of the ToF sensor. Other notions are introduced in the text.

We compare the depth in the *world CS*. Therefore, it requires two transformations to bring triangulated 3D points from stereo and the 3D points read out from the ToF sensor to this space. We use transformation matrix T_{sw} to transform the triangulated 3D points from the *stereo CS* to the *world CS*. We use transformation matrix T_{tw} to transform the 3D points from the *ToF CS* to the *world CS*.

We now introduce our method on how to compute T_{sw} and T_{tw}. We first calibrate stereo cameras and the ToF sensor, using traditional calibration methods described in [32], and then T_{sw} is computed as

$$T_{sw} = [R_1 \quad T_1], \tag{1}$$

where R_1 and T_1 are extrinsic parameters of the ToF camera. R_1 is a 3×3 rotation matrix and T_1 is a 3×1 translation vector.

T_{tw} cannot be *explicitly* defined because we do not know the relationship between the *ToF CS* and the *world CS* 2. Since the transformation of 3D points between these two spaces is rigid, we estimate its scale s, rotation R_2, and transformation T_2 to align these two coordinate systems

$$T_{rigid} = s * [R_2 \quad T_2]. \tag{2}$$

To reduce the errors involved in this process, we compute T_{tw} in sampled distance (we have a total of 16 sampled distances). In our experiments, we use a checkerboard that is movable on a metric board with two guide rails (Fig. 1). We manually select the corner points in the scene and read out X_{ToF} directly from SR3000. To compute its corresponding reference 3D points in the *world CS*, X_{world}, we first select these corners on the intensity image returned by the ToF sensor and find their corresponding corners in stereo cameras by using two homography matrices H_l and H_r. Because the patterns are planar, a homography matrix correctly describes the geometric

relationship from one view to another, and this matrix can be computed using at least four correspondences. In our experiments, we use the RANSAC [33] method to robustly estimate them. Since the projection matrices of the left and right stereo cameras are precalibrated, we can triangulate these corner features into stereo CS, and then transform them to world CS using T_{sw}.

Given X_{world} and X_{stereo}, we estimate T_{tw} using a closed form method [34]. We summarize the approach of computing T_{tw} in Algorithm 1.

Algorithm 1. Compute T_{tw}

1) Select corner pixels on the pattern, and acquire pixel correspondences from the ToF camera and stereo cameras, using H_l and H_r.
2) Triangulate the matched pixels to get 3D point X_{stereo}.
3) Transform them using T_{sw} to X_{world}.
4) Read out selected corner pixels' corresponding 3D points X_{ToF} from the ToF sensor.
5) Compute the rigid transformation matrix T_{tw} from 3D-3D correspondences (X_{world} and X_{ToF}) for each sampled distance.

Given T_{sw}, T_{tw}, we can compare depth difference for each pixel using

$$\Delta d(p) = T_{sw}X_{stereo} - T_{tw}X_{ToF}. \quad (3)$$

Δd is denoted as the geometric depth bias and stored in a LUT. In each sampled distance, this depth bias is calculated and added to the LUT. Finally, we have a 3D table whose cell stores the correction value $\Delta d(p)$.

4.3 Depth Refinement

Given LUTs from Photometric Calibration (LUT_p) and Geometric Calibration (LUT_g), the ToF sensor's depth measurements are refined as described in Algorithm 2.

Algorithm 2. Refine depth using LUTs from the ToF sensor.

1) Locate two cells in LUT_p, given the raw depth value D at pixel (u, v).
2) Compute depth bias $D_{bias} = a \times I + b$, where I is the intensity value at pixel (u, v), and a, b are parameters interpolated from values restored in LUT_p.
3) Compute refined depth using $D_1 = D + D_{bias}$, and transform D_1 to world CS using T_{tw}; this depth is denoted as D_2.
4) Locate two cells in LUT_g based on D_2 at pixel (u, v).
5) Linear interpolate depth bias D_{bias} from the bias values in these two cells.
6) Compute refined depth $D_3 = D_2 + D_{bias}$ in world space. D_3 is the final depth.

5 Calibration Verification

To evaluate the calibration result, we perform two experiments. The metrics we used in the evaluation are depth error (compared with ground truth depth) and Projection Residual Error (PRE). We first define these two metrics.

- **Depth error** is defined as the average distance between the reference depth (generated by passive stereo, using manually matched pixels) and any depth (returned from the ToF sensor or refined by the calibration results).
- **Projection Residual Error (PRE)** refers to the average difference between the 2D position of reference pixels (manually selected) and 3D points (returned from the ToF sensor or refined by the calibration results) that are projected back to the stereo cameras. We treat the Z value of the 3D points as the depth value.

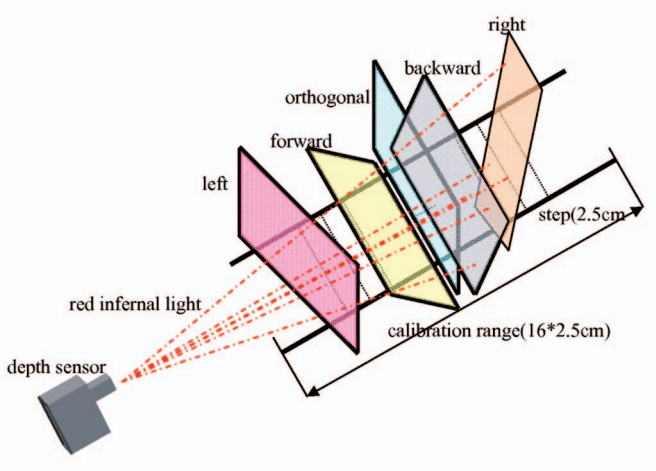


Fig. 5. LUT verification using planar plane. Five postures are tested, and each is placed at a different position inside the calibrated volume.

5.1 The Plane Experiment

In this experiment, a planar board with a checkered pattern is placed inside the calibrated volume. We test five postures: orthogonal, tilting forward, tilting backward, rotating left, and rotating right (see Fig. 5). Tilting and rotating are around 20 degrees (a rough estimation).

We show results from one of the postures in Fig. 6. We can see that both the 3D locations and the PRE from LUT-refinement results are much closer to the ground truth. Numerically, the depth error from selected samples of rigid transformation (after T_{tw}) is around 20 mm, while that is only 4 mm after LUT correction. We can also see the PRE from T_{tw} is almost 30 percent larger than that of LUT correction.

All of the numerical results of these five tests are presented in Table 1. Numbers in bold are after LUT refinement. PRE is small for the ToF sensor because of low resolution. This table also shows points after LUT-refinement have much smaller error, which is approximately one third of that original value.

5.2 The Box Experiment

In this experiment, we place a box inside the calibration volume and reconstruct the orthogonal patches. Our motivation is to show that the orthogonal relationship is better preserved by our LUT-refinement than the rigid transformation approach.

Similar to the plane test, we read out 3D points from the ToF sensor, compute T_{tw} transformed points, and refine them using LUT. In Fig. 7, we compare the angles returned by each method and reproject them into the left view. The angle between two patches is estimated, using a plane

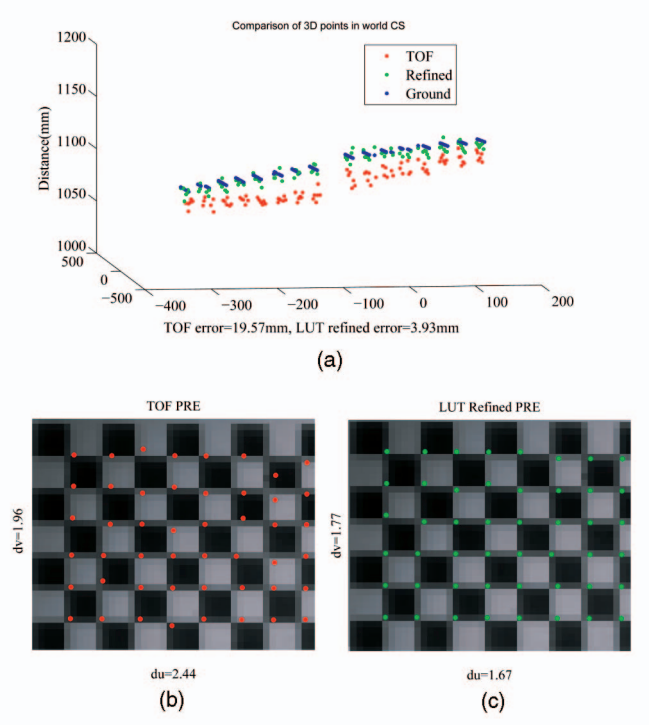


Fig. 6. Qualitative comparison of 3D points and PRE from the one of the planar tests. The ToF depth error is the average depth difference of the ToF data transformed by T_{tw}. LUT refined error is similarly computed after using photometric and geometric LUTs. The unit of PRE is pixel. (a) Comparison of 3D points in world CS. (b) ToF PRE. (c) LUT refinement PRE.

fitting approach. We can see that our LUT-refinement method preserves the angle well.

Based on the above tests, we verify that our calibration LUT improves the depth accuracy. To further improve the reconstruction quality, we present our fusion technique in the next section.

6 Sensor Fusion

The current state of the art in stereo matching is achieved by global optimization algorithms (e.g., [35], [36], [37]). These methods formulate stereo matching as a maximum a posterior Markov Random Fields (MAP-MRF) problem. In detail, we denote $X = x_i$ as hidden variables, corresponding to the disparities of each pixel and $Y = y_i$ as observed variables, corresponding to the intensity-based matching cost at specific disparity. Solving the stereo matching problem is equivalent to maximize the following posterior

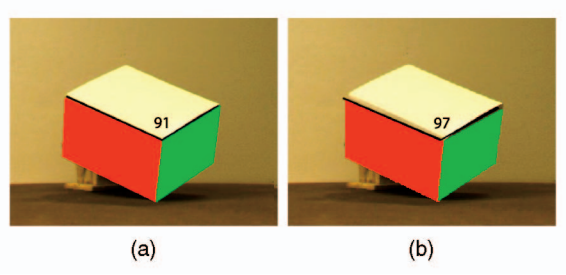


Fig. 7. Results from orthogonal patches. The angle from LUT-refinement is 91 degrees. The angle from T_{tw} transformation is 97 degrees. We can also see that the overlap of the fitted depth plane after using LUT-refinement is closer to the real orthogonal angle in the scene.

$$P(X|Y) \propto \prod_i f_d(x_i, y_i) \prod_{j \in N(i)} f_s(x_i, x_j), \tag{4}$$

where $N(i)$ represents the neighbors of node i, function f_d is the local evidence for node i based on the initial pixel-wise matching cost (*data term*), and f_s is a symmetric function that measures the smoothness assumption about the scene (*smoothness term*).

One valuable feature of this MAP-MRF formulation is that it provides a natural way to integrate the information from multiple sensors. With our ToF sensor, we add another set of observed variables $Z = z_i$, which corresponds to the depth value returned by the sensor. The new posterior can be formulated as

$$P(X|Y, Z) \propto \prod_i f_d(x_i, y_i) f_r(x_i, z_i) \prod_{j \in N(i)} f_s(x_i, x_j), \tag{5}$$

where $f_r(x_i, z_i)$ is the additional local evidence based on the measurement from the ToF sensor.

We choose Loopy Belief Propagation (LBP) to maximize the negative log-likelihood of $P(X|Y, Z))$. We also introduce two weighting factors to allow more flexibility for the data term. That is,

$$w_d \cdot \log f_d(x_i, y_i) + w_r \cdot \log f_r(x_i, z_i), \tag{6}$$

where w_d and w_r are the weighting factors for stereo and ToF data terms.

6.1 Configure MAP-MRF to Infer Depth

6.1.1 Data Term from Stereo Matching

The data term derived from stereo matching encodes the color consistency of pixel correspondences. In our implementation, pixel-wise matching costs are obtained in a similar manner as in [39]. In detail, the per-pixel difference is first computed, using Birchfield and Tomasi's pixel dissimilarity

TABLE 1
Result of Depth Error and Reprojection Error

Posture	Depth Error (mm)	PRE Sensor View (pixel)	PRE Left View (pixel)	PRE Right View (pixel)
orthogonal	17.0 **(5.3)**	(0.34 0.36) **(0.02 0.04)**	(1.82 2.35) **(1.85 1.63)**	(3.50 2.21) **(1.91 1.72)**
left	17.0 **(6.1)**	(0.35 0.28) **(0.03 0.06)**	(3.33 2.68) **(1.87 1.64)**	(1.68 2.62) **(2.44 1.65)**
right	20.3 **(5.0)**	(0.32 0.35) **(0.02 0.05)**	(1.96 2.44) **(1.77 1.67)**	(4.21 2.23) **(2.91 1.65)**
forward	19.2 **(5.1)**	(0.39 0.31) **(0.02 0.08)**	(6.24 4.72) **(3.75 3.33)**	(5.81 4.60) **(4.41 3.14)**
backward	16.3 **(4.6)**	(0.36 0.33) **(0.02 0.06)**	(2.44 2.48) **(1.86 1.69)**	(3.12 2.31) **(2.10 1.71)**

[40]. An additional adaptive weight aggregation step [41] is applied to overcome matching ambiguities caused by occlusion boundaries or sensor noise. This two-step approach has been shown to be remarkably effective for getting reliable matching cost in [39].

6.1.2 Data Term from the ToF Sensor

The data term from the ToF sensor encodes the depth consistency between the stereo and the ToF sensor in *world CS*. We use the same technique to compute this term as our previous work in [42].

We describe the process of computing the ToF term briefly. From the stereo cameras, for each pixel p in the left view, we have a set of disparity candidates d_c. By triangulating each matched pair, we get a set of 3D points in the *stereo CS*. Then we transform them, using T_{sw} to the *world CS*. From the ToF sensor, we read the 3D points and also transform them to the *world CS*, using T_{tw}.

By assuming that the true 3D point is one of the points from the stereo triangulation, we define the cost of the ToF term as the difference between these two types' 3D points

$$f_r = \min(|x_{stereo} - x_{ToF}|, \tau_1), \tag{7}$$

where τ_1 is set to 300 mm. X_{stereo} and X_{ToF} are both transformed to *world CS*, using T_{sw}, T_{tw}, respectively. Note that in [38], we used an exponential function to make the cost smoother. We found that a simple linear function also works well.

6.1.3 Smoothness Term

The smoothness term encodes a prior assumption that depth should be piecewise smooth. We use a quadratic truncation model to describe this term

$$f_s = \min\{(x_i - x_j)^2, \tau_2\}, j \in N_s(i), \tag{8}$$

where x_i and x_j are the disparity values of selected pixel i and its neighboring pixels j; τ_2 is set to half of the maximum disparity value.

The occlusion between left and right cameras may cause visible problems on the boundaries in the disparity map. We employ a similar approach as [36], where the occlusion map is estimated by minimizing an MRF-based cost function for measuring the consistence of occlusion between left and right cameras.

7 Fusion by Reliability

The weighting functions w_d and w_r in (6) have to be carefully selected. This is difficult in practice because both stereo matching and measurements from the ToF sensor may not be robust. In this section, we introduce our per-pixel reliability function that can aid in computing w_d and w_r adaptively instead of tuning the parameters as in our previous work [42]. The per-pixel reliability is represented by a coefficient that *weights* the per-pixel cost. Results show fusion by reliability can correctly describe the unambiguous characteristic for each method and yield better quality.

7.1 Reliability of Passive Depth

The definition of reliability in stereo matching is simple: The best depth candidate should have a distinct matching score. We intuitively define the matching reliability of pixel p as how distinctive the best and the second best costs are:

$$R_d(p) = \begin{cases} 1 - \frac{c_p^{1st}}{c_p^{2nd}} & C_p^{2nd} > T_c \\ 0 & otherwise, \end{cases} \tag{9}$$

where c_p^{1st} and c_p^{2nd} are the best (lowest) and the second best matching costs of p. T_c is a small threshold to avoid division by zeros. $R_d(p) \in [0, 1]$ can be used to model ambiguous matching by poor SNR [43].

7.2 Reliability of Active Depth

Our motivation for computing the reliability of the ToF sensor is to incorporate its robustness in the fusion framework. We can use a similar approach as to the first and second costs to define the reliability of the ToF sensor as in our previous experiment [38]. However, in this paper, we would like to explore the essential reliability of the ToF sensor.

We first introduce our result that the the reliability of depth measurements from the ToF sensor depends on the lightness that the ToF sensor can receive (we use the returned amplitude from the ToF sensor to measure the lightness). We then discuss the two main factors that cause fall off in lightness: object reflectance and vignetting, followed by the introduction to our method to model the reliability of the ToF sensor.

Object Reflectance. As we already demonstrated in the calibration section, darker objects tend to absorb photons, which reduce the amplitude and the brightness the ToF sensor returned.

Vignetting. In addition to object reflectance, vignetting also attenuates the brightness, particularly near the image edge, and affects the accuracy of depth estimation [44]. Several previous works in the literature have also reported that circular errors from the imperfect sinusoidal modulation of the transmitted near infrared light [15], [25] cause lightness loss from the image center to the image edge.

From a real scene, object reflectance and vignetting are difficult to measure independently. Therefore, directly biased evaluation of depth estimation can be problematic. In this paper, we provide an approach by modeling a reliability function using amplitude, which is a good metric to measure how much light is returned.

To fit the reliability function, we do a statistical analysis of how object reflectance and vignetting can change the amplitude. We set the integration time to 40, and for each experiment, we collect 1,000 measurements from the ToF sensor.

Verification. We first place a white board in the scene. Fig. 8a shows the amplitude. We downsample the original 16-bit amplitude, using an intensity image which is 8-bit. The further an object is from the center, the greater the reduction in brightness. We also plot the per-pixel standard deviation (PPSD) (in color plot) of the depth. From the PPSD, we find that *the further from the center, the higher the deviation is*, which shows that the accuracy of depth measurement from the ToF sensor decreases with the falloff of the amplitude.

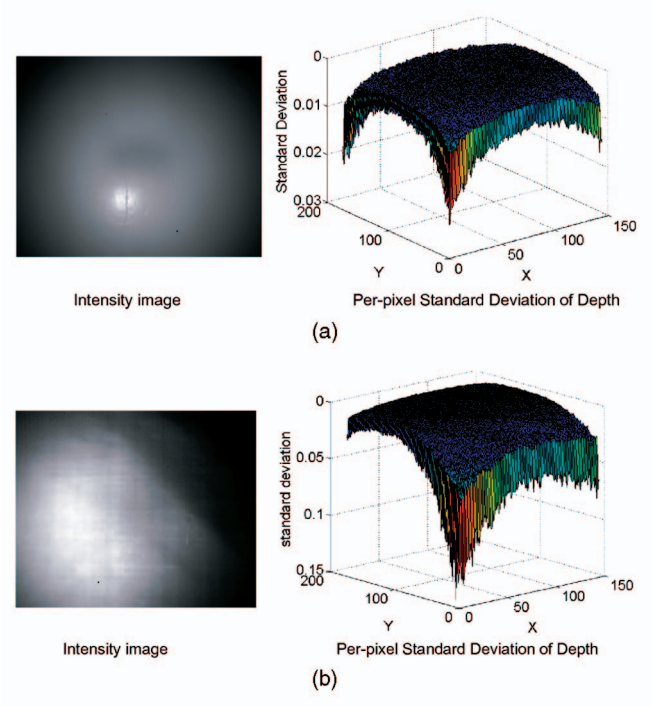


Fig. 8. Example of vignetting and object reflectance from a planar object. Vignetting and object reflectance tests from a planar object. Images in (a) and (b) denote the amplitude returned from the ToF sensor. The image in (a) is captured while the ToF sensor faces a white planar board. We can see the vignetting effects with darker pixels near the image boundary. The image in (b) is captured while the ToF sensor faces a planar board with intensity values reduced from left to right. We can see the image in (b) shows a combination of vignetting and intensity changing effects. We also plot the standard deviation from 1,000 measurements in the color plot for both tests. We can see the std is larger when the amplitude returned is less. (a) Object in white. (b) Object with various reflectance.

In the second experiment, we measure the PPSD from a planar object with various reflectances (a planar board painted with gradually changed intensity values). In Fig. 8b, one can see small PPSDs for higher object reflectance and large PPSDs for lower object reflectance.

Statistical Analysis. We sample many different object reflectances from a number of scenes and collect all of the PPSDs, which are shown in Fig. 9a. The PPSDs are normalized into $[0, 1]$. We can see the standard deviation of depth measurement is inversely proportional to the amplitude. Based on this observation, we assume that amplitude is proportional to the ToF sensor's reliability and we calibrate the reliability using the inverse proportional to the standard deviation of the amplitude.

We collect all possible amplitudes and divide them into 256 bins. We then fit a normal distribution to the depth deviations in each bin, and the inverse of the normalized standard deviation is used to model the reliability:[1]

$$R_r(p) = 1/\sigma, \tag{10}$$

where p denotes the bin that a specified pixel is located based on its amplitude. σ is the fitted standard deviation of amplitudes from a normal distribution with 0.95 percent confidence intervals.

1. Note that the numerical results in Table 1 were obtained with a method differing slightly from (10) (originally, we used an exponential function to smooth the reliability) and (10) was observed to give almost the same results but in a simple way.

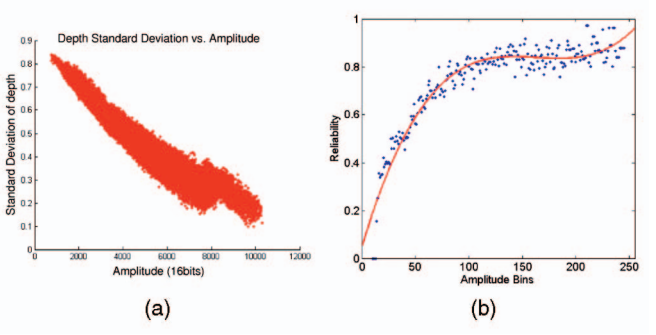


Fig. 9. Our reliability function based on amplitudes. (a) shows the relationship between amplitudes and standard deviation of depth measurement. (b) shows the experimental curve model used in this paper. (a) Standard deviation of depth measurement. (b) Reliability model using curve function.

In practice, we are not able to sample all of the variations as variations such as amplitude values are very small or high. Instead, we plot all available amplitudes and fit them, using a curve function. We found that a cubic curve (plotted in blue in Fig. 9b) can experimentally approximate the reliability samples well.

Given the reliability curve function, we compute the per-pixel reliability $R_r(p)$ for the depth measurements from the ToF sensor. Finally, given $R_d(p)$ and $R_r(p)$, we compute $w_d = \frac{R_d(p_i)}{R_d(p_i)+R_r(p_i)}$ and $w_r = 1 - w_d$.

8 Experimental Results

In this section, we evaluate the quality of depth estimation from traditional stereo matching, the ToF sensor, and our fusion approach. In all experiments, we use the left view of the stereo cameras as the reference view.

In order to verify the accuracy of these three methods, we set up a single structured light scanner to acquire reference scene depth. Basically, we use a projector to project a single line sweeping over the scene. The orientation of the line is roughly orthogonal to the epipolar line of the cameras. Therefore, the correspondence problem can be uniquely determined, using the stereo images. Based on the cameras' calibration data and projector resolution ($1{,}024 \times 768$, we sweep 1,024 lines over the scene), our structured light scanner is able to achieve very high accuracy depth maps, which we use as the ground truth to obtain quantitative results.

We compare our results in three groups: depth maps from passive stereo and those from the ToF sensor alone, depth maps from local fusion (without **MAP-MRF**) and global fusion (with **MAP-MRF**), and depth maps from non-reliability fusion and reliability fusion. In the first set of comparisons, we demonstrate that depth from passive stereo and that from active sensors have complementary characteristics in nature; the second set of comparisons shows that our global fusion approach can substantially increase the overall depth accuracy by a three times error reduction; in the third set of comparisons, we show that our reliability approach can infer the depth details better.

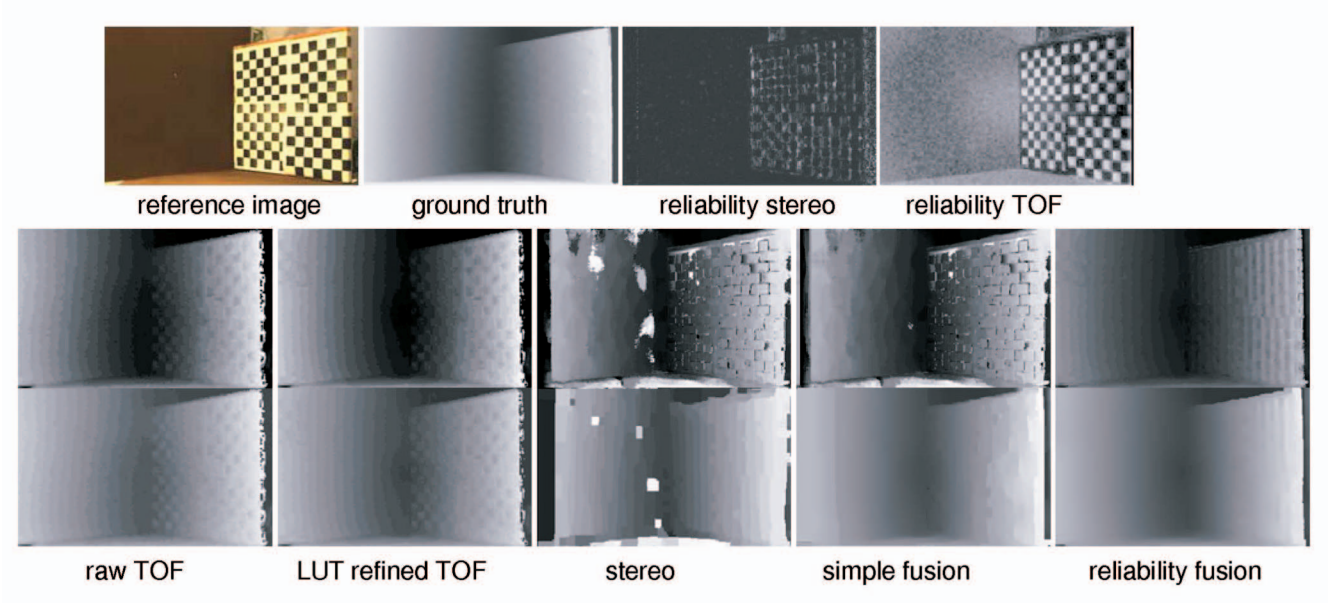


Fig. 10. Depth map from a simple scene with two planetary boards. The first row shows the reference image, our ground truth, and two maps of per-pixel reliability from passive stereo and the ToF sensor. The second row shows the depth map from local methods. The third row shows the depth map from global methods. From left to right are: raw depth from the ToF sensor, LUT-refinement depth from the ToF sensor, depth from stereo matching, depth from S-Fusion, and depth from R-Fusion.

8.1 Definition of Methods

We first introduce the different approaches used in our experiments.

The distinction of local and global methods originally stems from stereo vision literature. As defined in the taxonomy by Scharstein and Szeliski [5], a local method associates a pixel's disparity value to the one with the minimum matching cost, e.g., a local "winner-takes-all" (WTA) approach. In contrast, a global method typically makes disparity decision using an energy-minimization framework, which is formulated as an **MAP-MRF** solution (see Section 6) in this paper.

We define *local methods* to compute the depth maps from passive stereo and the ToF sensor alone. The former method is a two-step approach by: 1) calculating the pixel similarity, and 2) applying an adaptive weight aggregation. The second method essentially transforms the raw depth values returned from the ToF sensor to the *stereo CS* and computes the disparities.

By setting w_d and w_r in (6) to different values, we define the second type of comparison methods. By setting w_r to zero, we obtain global depth results from passive stereo; by setting w_d to zero, we obtain global depth results from the ToF sensor. We also define a *local fusion method* by setting $w_d = w_r = 0.5$, and a *global fusion method* which applies **MAP-MRF** from the *local fusion method*.

Finally, we define the *reliability local fusion method* and *reliability global fusion method* by setting w_d and w_r, using (9) and (10), which essentially incorporate a per-pixel weighting function.

In abbreviation, we refer to the method using $w_d = w_r = 0.5$ to as Simple Fusion (*S-Fusion*), the one using per-pixel reliability as Reliability Fusion (*R-Fusion*). This is also the method we expected to obtain the highest accuracy.

8.2 Results on Simple Scenes

We evaluate all methods on a number of scenes. For each scene, we first use the scanner to obtain the ground truth, then apply different methods to compute the depth maps. We first present results from a simple scene with two planes (scene Plane): One is uniformly colored, the other is a checkerboard.

As expected (see Fig. 10), the ToF sensor and passive stereo exhibit complementary characteristics; stereo matching is erroneous in weakly/periodic textured areas while the ToF sensor obtains unstable depth on boundaries and dark regions. This unambiguous characteristic is also present in the reliability maps; the reliability on the textureless plane (left) from the ToF sensor is higher overall than that from passive stereo; and passive stereo has a greater reliability on the rich-texture region (right).

It is also interesting to observe that the sensor's depth maps are (in the first and second row) improved after applying a global optimization step. It is expected that the depth map from R-Fusion is preferable and more visually pleasing than that from S-Fusion, especially on textured regions and around occlusion boundaries. In contrast, as the reliability map provides knowledge on how to *intelligently* weight different nodes in the MRF graph, their information can be more properly propagated.

It is also notable that global methods overwhelm local methods in our experiments. Therefore, for the following experiments, we present only global results.

8.3 Results on Complex Scenes

In this section, we present results from two complex scenes (scene Teddy and scene Head). Depth inference for these scenes is challenging because of occlusions and thin structures. We demonstrate results from the global method in Fig. 11.

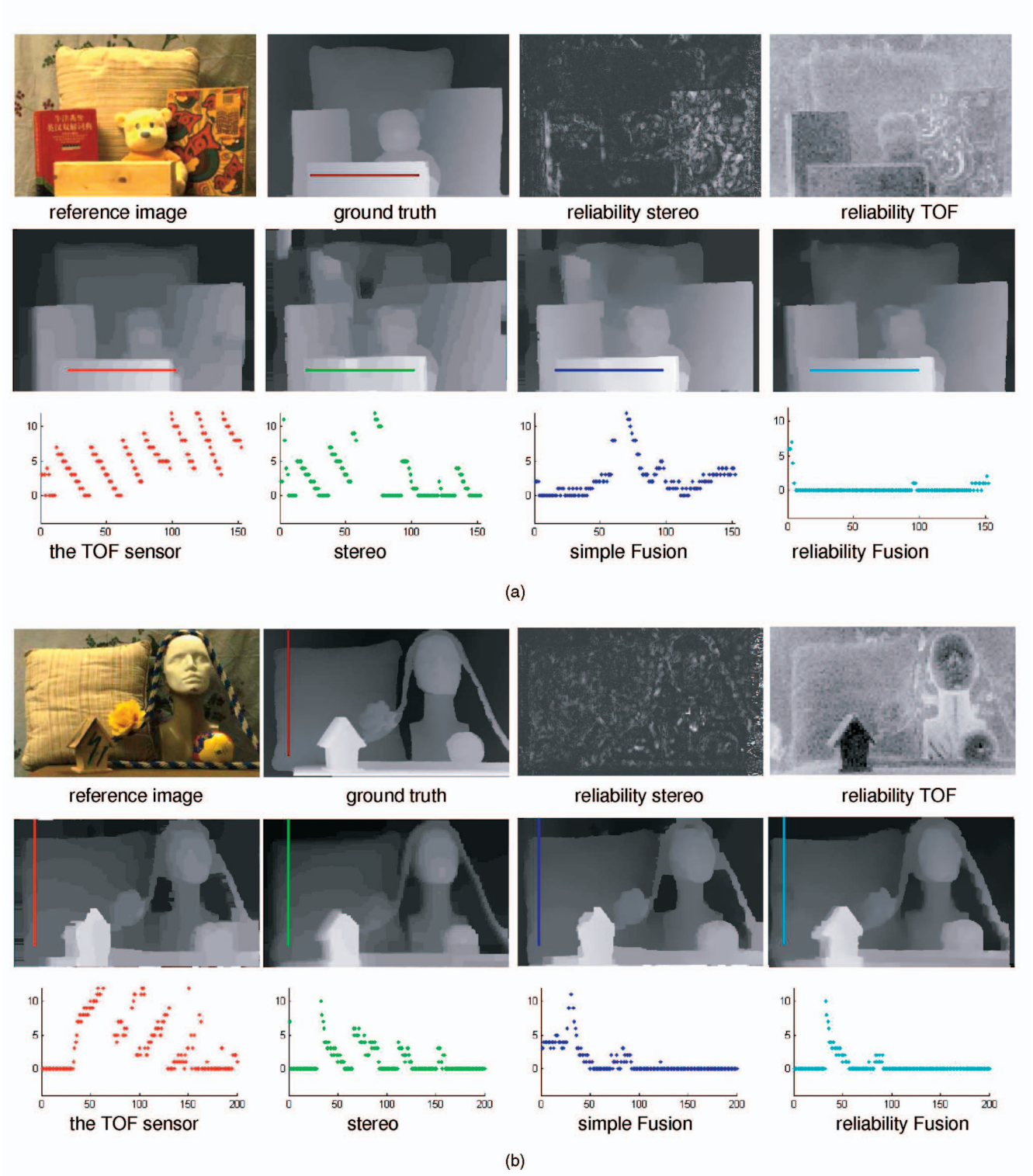


Fig. 11. Results from two complex scenes. The first row shows the reference image, our ground truth, and two maps of per-pixel reliability from passive stereo and the ToF sensor. The second rows shows depth maps from global methods. From left to right: raw depth from the ToF sensor, LUT-refinement depth from the ToF sensor, depth from stereo matching, depth from S-Fusion, and depth from R-Fusion. The third row shows the gradient of depth samples from one row/column. (a) Scene Teddy. (b) Scene Head.

From the reliability maps, we can see the overall complementary nature is presented (compare reliability of pillow in both scenes from both methods). This is consistent with our assumption. However, the reliability from the ToF sensor is notable, as it varies with different object materials. In scene Head, the head region is brighter than the neck region in the real scene; however, their reliability exhibits contrary behavior.

We point out that the high intensity in the color image does not represent high amplitude the ToF sensor returned.

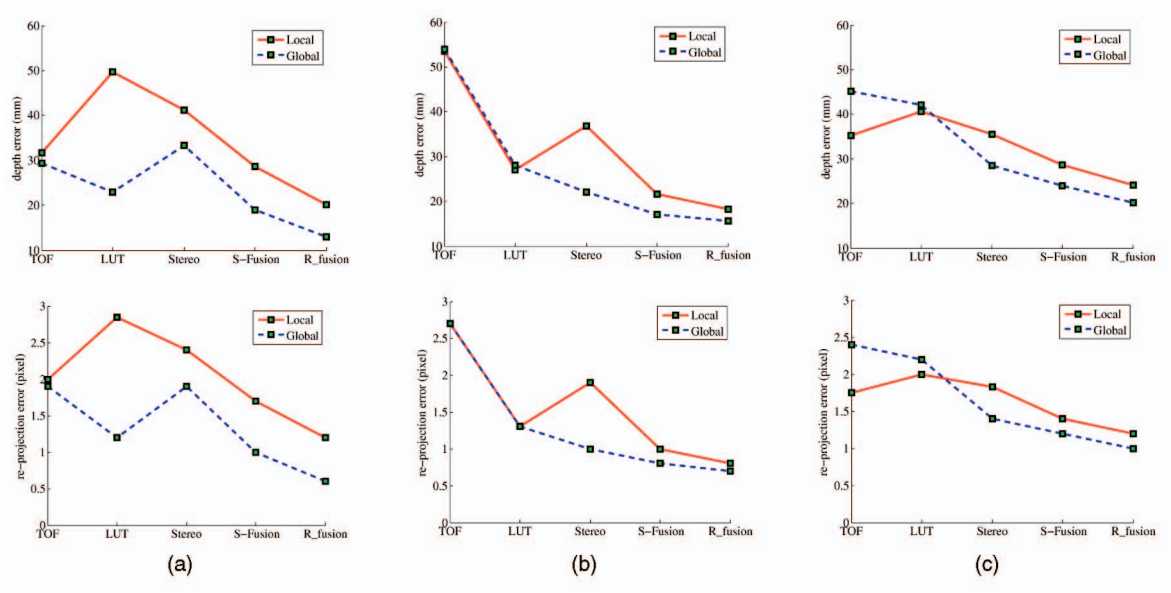


Fig. 12. Comparison of all methods with ground truth. From left to right: Scenes Plane, Teddy and Head. The first row compares the depth error in mm, the second row compares the reprojection error in pixels. (a) Scene Plane. (b) Scene Teddy. (c) Scene Head.

We noted that in this case, the amplitude in the neck region is actually higher than the face region since they are captured under different lighting conditions. The ToF sensor uses active IR lights, while the color image is captured with an IR blocking filter. In addition, we believe our reliability function is not perfect because the high reflectance causes unknown behavior, for example, the neck/head case and the cup case we discussed in the paper. We also think this is a field that requires further research.

Other interesting reliability behaviors of the ToF sensor are: The reliability of the wood in scene Teddy is not uniform; the reliability of the small house in scene Head changes rapidly (compare the roof and the front side). Reasons that may account for this are the complex light redirection and object material. We believe the ToF sensors in these fields are yet to be explored.

Nevertheless, the improvement from R-Fusion is promising compared to other methods, such as the regions near the pillow and the book in scene Teddy and the small house in scene Head. In addition, R-Fusion provides smoother depth maps, as we note that the global method using either stereo or the ToF sensor alone results in incorrect discontinuities; S-Fusion reduces the depth map's noise and provides better results on depth discontinuities. However, it cannot preserve all of the details, while our R-Fusion demonstrates better performance.

8.4 Numerical Comparison

The numerical comparison against the ground truth from these three scenes is presented in Fig. 12 and Table 2. We compare the mean depth error (in millimeters) and the mean disparity error (in pixels, the number in *italic*). We summarize the results as the following:

1. Both local and global results by LUT are better than raw sensor output (directly from the ToF sensor). This again reflects the importance of our calibration process.
2. Stereo matching does poorly on textureless or repetitive regions. This can be verified by the error from scene Plane, which is larger than that of complex scenes.
3. By comparing numbers in columns of local and global method, we can see the global method works well on simple scenes.
4. S-Fusion can reduce the depth error by more than half, on average, as compared to the raw sensor data (after rigid transformation).
5. R-Fusion obtains the best result and can reduce the error by almost 20 percent compared to S-Fusion.

By comparing Table 2 with Table 1, we can see that the reconstruction accuracy of complex scenes is not as useful as that of the simple scenes. We believe it is due to complex lighting, surface reflectance, and texture variations.

8.5 Special Scenes

In this section, we first discuss results from two special scenes (high reflectance and transparent object). We show that the fusion approach is less robust in such cases. Second, we apply our approach to high-speed cameras which return only gray scale images. This decreases the quality of the stereo results because color aggregation is not available. Nevertheless, we show that our fusion approach can still achieve acceptable results.

8.6 High Reflectance and Transparency

We evaluate global methods with two additional scenes (scene Cup and scene Book). These two scenes are special because the former has strong specular/inter-reflection and the latter has transparent materials. Both of them can distort the result from a structured light scanner; therefore, they are not included for numerical evaluation.

We first show the results from scene Cup in Fig. 13. We can see that the depth map from structured light method is completely inaccurate on some scan lines. This is because

TABLE 2
Numerical Comparison of Real Scenes

	scene: Plane		scene: Teddy		scene: Head	
	Local	Global	Local	Global	Local	Global
raw sensor	31.7 (*2.0*)	29.3 (*1.9*)	53.5 (*2.7*)	53.97 (*2.7*)	35.2 (*1.75*)	45.1 (*2.4*)
LUT-refinement	49.7 (*2.85*)	23.0 (*1.2*)	27.0 (*1.3*)	28.8 (*1.3*)	40.6 (*2.0*)	42.1 (*2.2*)
stereo	41.2 (*2.4*)	33.3 (*1.9*)	36.8 (*1.9*)	22.0 (*1.0*)	35.5 (*1.83*)	28.5 (*1.4*)
S-Fusion	28.6 (*1.7*)	19.0 (*1.0*)	21.6 (*1.0*)	17.1 (*0.8*)	28.7 (*1.4*)	24.0 (*1.2*)
R-Fusion	20.1 (*1.2*)	13.0 (*0.6*)	18.3 (*0.8*)	15.6 (*0.7*)	24.1 (*1.2*)	20.2 (*1.0*)

the dark color projected on the high specular reflection spot appears to be white, resulting in severe matching errors. The reliability map from the ToF sensor is not correct in regions with specular refection presented (see the cup and the bottle). The output from the ToF sensor are grossly wrong on the inter-reflection region (right side of the cup in scene Cup). It is surprising that the global stereo matching method, which assumes a Lambertian scene, is in fact quite robust on objects with small specular highlights or inter-reflections, while both structured light and the active method failed. Given the inaccurate confidence map from the ToF sensor, R-Fusion does poorly on estimating the depth for the cup and the bottle, while S-Fusion can still obtain acceptable results.

In Fig. 14, we show the results from scene Book. The depth reported by the ToF sensor is completely inaccurate on transparent materials. Due to a severe matching error, stereo matching cannot generate good results. Therefore, both S-Fusion and R-Fusion failed in this case as the cup is transparent. Nevertheless, compared with S-Fusion, the R-Fusion method is better as it returns the shape of the cup by suppressing the ToF sensor more on transparent regions.

From these two experiments, we observe that the ToF sensor is very susceptible to high specularity, inter-reflection, and transparency. This is a direction in which sensor fusion can be further explored.

8.7 Extended Calibration Volume, High Frequency Cameras, Grayscale Images and Large Baseline Stereo

In this section, we present supplementary tests on an extended calibration volume. The volume is extended to around 3 m, and we employ the same calibration method introduced in Section 4 to calibrate the system. We use high frequency cameras that can only return monochromatic images. We double the baseline of the stereo in one of the tests.

Our setup in Section 3 uses the cameras with capture frequency at a speed of 15 FPS, which can return very decent color images in $1{,}024 \times 768$ resolution. Increasing the capture frequency is useful, as the capture frequency of the ToF sensor (SR4000, the new version of Swissranger, reaches 56 FPS at most) is faster than the cameras used in our test. However, by increasing the capture frequency, the quality of images is decreased (such as color information is lost). We present results on low quality color images from

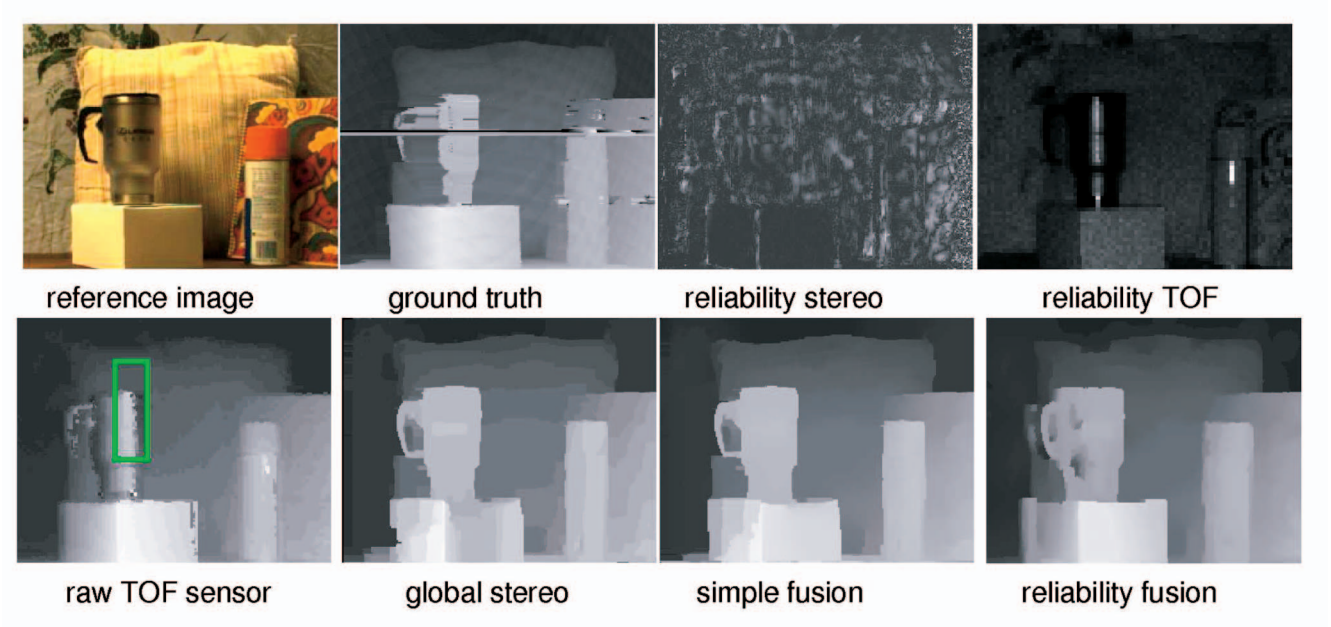


Fig. 13. Results from scene Cup. Results show that the ToF sensor returns incorrect depth both from specular and inter-reflection regions. S-Fusion generates acceptable results, while R-Fusion does poorly because per-pixel reliability from the ToF sensor is incorrect.

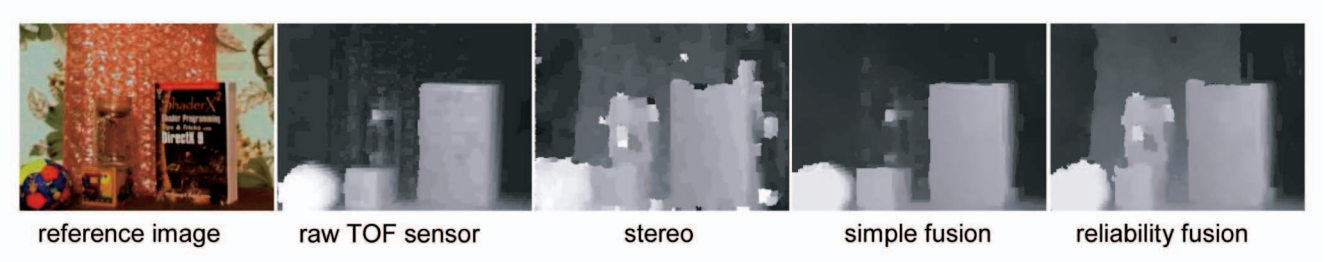


Fig. 14. Results from scene Book. None of the approaches can obtain acceptable depth results on the transparent materials (the cup).

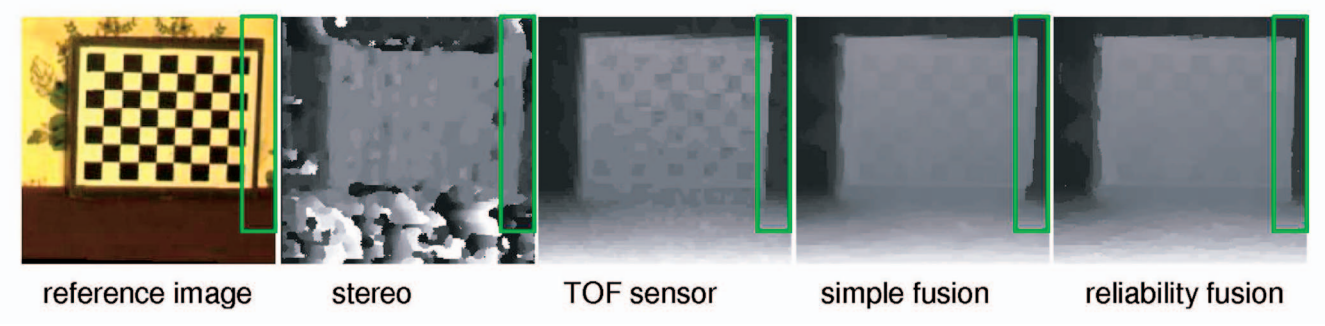


Fig. 15. Results from low quality color images. We can see that the disparity map from our fusion approach is better than stereo or the ToF sensor alone. Our R-Fusion achieves the best results by preserving details inside the green box.

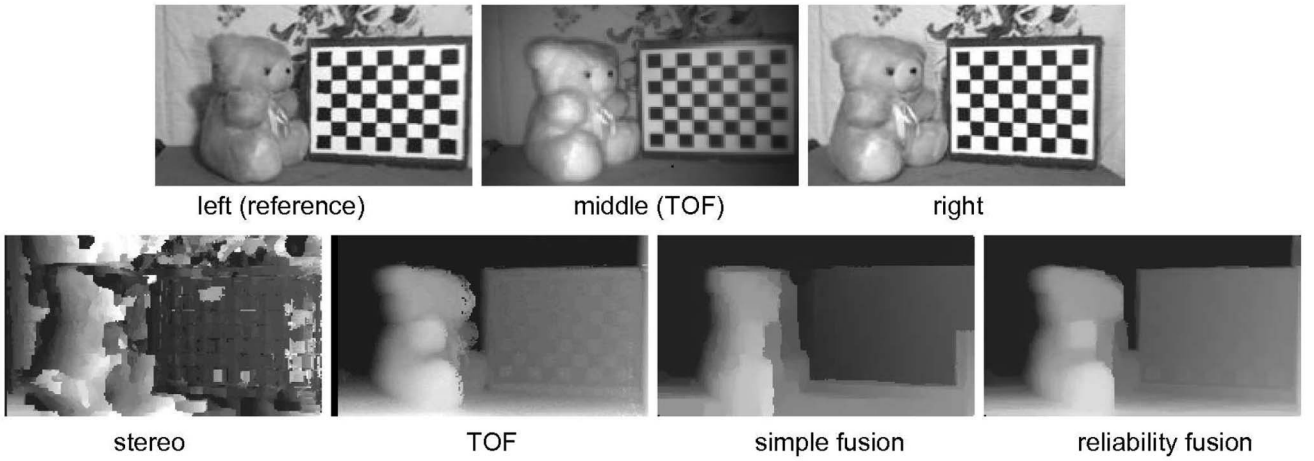


Fig. 16. Results from large baseline and high-speed cameras. The distance of objects from the setup is around 2.0 m. The first row shows the left, ToF, and right view of the. scene. The baseline between the two cameras is around 20 cm (original setup is around 10 cm). The capturing speed of stereo cameras is around 100 FPS. The capturing speed of the ToF sensor is around 12 FPS as we set the integration time to 50.

two high-speed cameras (200 FPS with image resolution of 640×480, as we introduced in our previous work [38]).

We display an example and compare results in Fig. 15. The scene is very simple and contains only one checkerboard pattern. By comparing all of the depth maps from the right edge on the checkerboard, we see the boundary of the checker board from R-Fusion is more accurate because R-Fusion can assign large weights to the ToF sensor on the region below the green box.

We then show an example with a large baseline. The stereo cameras capture the scene at a high speed of 100 FPS. The returned images from stereo cameras are in grayscale. This makes our color aggregation step in stereo matching unavailable. Fig. 16 shows our results. We note that the overall depth map is correct. Due to a large occlusion and lack of color information, the depth from the stereo is very poor. By fusing with the ToF sensor, we are still able to obtain acceptable results.

9 Conclusion

In this paper, we present a practical calibration method to improve the performance of Time-of-Flight (ToF) sensors. Our method is general and needs no additional equipment other than a pair of cameras. We use the stereo camera to generate reference depth values in calibration so that the sensor can be calibrated for any desired range. Evaluation shows that our calibrated depth map can achieve an absolute accuracy of about 5 mm over the range of 1 m. This improvement is approximately three times better compared to the raw depth map (after rigid alignment).

We also present a fusion method that is useful in improving depth quality by maximizing complementary

information from passive and active depths. The nature of our approach is based on an accurate per-pixel reliability calculation for both methods. We show comprehensive results from different scenes and compare them with state-of-the-art depth estimation methods such as structured light method and stereo. Results show that our S-Fusion reduces the overall error by 50 percent, and our R-Fusion can further reduce this error by almost 20 percent. Nevertheless, there are also problematic cases, such as high reflectance and transparency. For such cases, the complementary nature of the ToF sensor and stereo is invalid; therefore, our formulation cannot deal with these problems.

Our fusion approach requires around 20 s to generate a depth map in 400×300 resolution. The main computation power is used for LBP. Although it is currently not in real time, several acceleration approaches of LBP using GPU [45], [46], [47] are already available. We envision extending our method to a hardware-accelerated approach in the near future.

ACKNOWLEDGMENTS

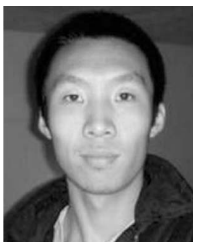
Ruigang Yang was supported by the University of Kentucky Research Foundation, the US Department of Homeland Security, US National Science Foundation (NSF) HCC-0448185, and CPA-0811647. James E. Davis was supported by NSF CCF-0746690. Zhigeng Pan was supported by the China NSFC 60533080 and the China 863 Plans 2006AA01Z335. The authors thank Qing Zhang and Xueqing Xiang for collecting part of the data. This work was done when Jiejie Zhu was with the University of Kentucky as a postdoctoral researcher.

REFERENCES

[1] "Canesta Inc., Canestavision Electronic Perception Development Kit," http://www.canesta.com/, 2006.
[2] "Swissranger Inc., sr-2," http://www.csem.ch/fs/imaging.htm, 2006.
[3] "3dv Systems, z-cam," http://www.3dvsystems.com, 2004.
[4] "Photonix Mixer Device for Distance Measurement," http://www.pmdtec.com, 2001.
[5] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int'l J. Computer Vision,* vol. 47, no. 1, pp. 7-42, 2002.
[6] A. Verri and V. Torre, "Absolute Depth Estimate in Stereopsis," *J. Optical Soc. of Am. A,* vol. 3, pp. 297-299, 1986.
[7] Z. Xu, R. Schwarte, H. Heinol, B. Buxbaum, and T. Ringbeck, "Smart Pixel-Hptonic Mixer Device (PMD)," *Proc. Int'l Conf. Mechatron and Machine Vision,* pp. 259-264, 1998.
[8] M. Lehmann, R. Kaufmann, F. Lustenberger, B. Büttgen, and T. Oggier, "CCD/CMOS Lock-In Pixel for Range Imaging: Challenges, Limitations and State-of-the-Art," CSEM, Swiss Center for Electronics and Microtechnology, 2004.
[9] T. Oggier, B. Büttgen, F. Lustenberger, G. Becker, B. Rüegg, and A. Hodac, "Swissranger SR3000 and First Experiences Based on Miniaturized 3D-ToF Cameras," *Proc. Conf. First Range Imaging Research Day,* 2005.
[10] G.J. Iddan and G. Yahav, "3D Imaging in the Studio," *Proc. SPIE Conf.,* pp. 48-56, 2001.
[11] R. Lange and P. Seitz, "Solid-State Time-of-Flight Range Camera," *IEEE J. Quantum Electronics,* vol. 37, no. 3, pp. 390-397, Mar. 2001.
[12] S.Y. Chen, *Active Sensor Planning for Multiview Vision Tasks.* Springer, 2008.
[13] S.Y. Chen, Y.F. Li, and J.W. Zhang, "Vision Processing for Realtime 3D Data Acquisition Based on Coded Structured Light," *IEEE Trans. Image Processing,* vol. 17, no. 2, pp. 167-176, Feb. 2008.
[14] S.A. Gudmundsson, H. Aanæs, and R. Larsen, "Environmental Effects on Measurement Uncertainties of Time-of-Flight Cameras," *Proc. Int'l Symp. Signals, Circuits, and Systems,* pp. 1-4, 2007.
[15] T. Kahlmann, F. Remondino, and H. Ingensand, "Calibration of the Fast Range Imaging Camera Swissranger for Use in the Surveillance of the Environment," *Proc. SPIE Conf. Electro-Optical Remote Sensing II,* 2006.
[16] H. Gonzales-Banos and J. Davis, "Computing Depth Under Ambient Illumination Using Multi-Shuttered Light," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recoginition,* pp. 234-241, 2004.
[17] M. Lindner and A. Kolb, "Calibration of the Intensity-Related Distance Error of the PMD ToF-Camera," *Proc. SPIE Conf. Intelligent Robots and Computer Vision XXV,* 2007.
[18] S. Fuchs and G. Hirzinger, "Extrinsic and Depth Calibration of ToF-Cameras," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 464-468, 2008.
[19] B. Huhle, T. Schairer, P. Jenke, and W. Straber, "Robust Non-Local Denoising of Colored Depth Data," *Proc. First Workshop Time-of-Flight Based Computer Vision,* 2008.
[20] S.B. Göktürk, H. Yalcin, and C. Bamji, "A Time-of-Flight Depth Sensor: System Description, Issues, and Solutions," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recoginition,* p. 35, 2004.
[21] Q.X. Yang, R.G. Yang, J. Davis, and D. Nister, "Spatial-Depth Super Resolution for Range Images," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2007.
[22] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "High-Quality Scanning Using Time-of-Flight Depth Superresolution," *Proc. First Workshop Time-of-Flight Based Computer Vision,* 2008.
[23] J. Diebel and S. Thrun, "An Application of Markov Random Fields to Range Sensing," *Advances in Neural Information Processing Systems,* pp. 291-298, MIT Press, 2005.
[24] A.H. Izhal, T. Ushinaga, T. Sawada, M. Homma, Y. Maeda, and S. Kawahito, "A Cmos Time-of-Flight Range Image Sensor with Gates-on-Field-oxide Structure," *IEEE Sensors J.,* vol. 7, no. 12, pp. 1578-1586, Dec. 2007.
[25] M. Lindner, A. Kolb, and K. Hartmann, "Data-Fusion of PMD-Based Distance-Information and High-Resolution RGB-Images," *Proc. Int'l Symp. Signals, Circuits, and Systems,* 2007.
[26] R. Reulke, "Combination of Distance Data with High Resolution Images," *Proc. Conf. Image Engineering and Vision Metrology,* 2006.
[27] T.D.A. Prasad, K. Hartmann, W. Weihs, S.E. Ghobadi, and A. Sluiter, "First Steps in Enhancing 3D Vision Technique Using 2D/3D Sensors," *Proc. Computer Vision Winter Workshop,* pp. 82-86, 2006.
[28] C. Beder, B. Bartczak, and R. Koch, "A Comparison of PMD-Cameras and Stereo-Vision for the Task of Surface Reconstruction Using Patchlets," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recoginition,* 2007.
[29] S.A. Gudmundsson, H. Aanæs, and R. Larsen, "Fusion of Stereo Vision and Time-of-Flight Imaging for Improved 3D Estimation," *Proc. Int'l Workshop Dynamic 3D Imaging,* 2007.
[30] S. May, B. Werner, H. Surmann, and K. Pervölz, "3D Time-of-Flight Cameras for Mobile Robotics," *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems,* pp. 1578-1586, 2006.
[31] L. Guan and M. Pollefeys, "A Unified Approach to Calibrate a Network of Camcorders and ToF Cameras," *Proc. IEEE Workshop Multi-Camera and Multi-Model Sensor Fusion Algorithms and Applications,* 2008.
[32] Z. Zhang, "A Flexible New Technique for Camera Calibration," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 11, pp. 1330-1334, Nov. 2000.
[33] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision,* second ed. Cambridge Univ. Press, 2003.
[34] B.K.P. Horn, "Closed-Form Solution of Absolute Orientation Using Unit Quaternions," *J. Optical Soc. of Am.,* vol. 4, no. 4, pp. 629-642, 1987.
[35] V. Kolmogorov and R. Zabih, "Computing Visual Correspondence with Occlusions Using Graph Cuts," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 508-515, 2001.
[36] J. Sun, Y. Liy, S.B. Kang, and H.Y. Shum, "Symmetric Stereo Matching for Occlusion Handling," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recoginition,* pp. 399-406, 2005.
[37] Q.X. Yang, L. Wang, R. G.Yang, S. G.Wang, M. Liao, and D. Nister, "Real-Time Global Stereo Matching Using Hierarchical Belief Propagation," *Proc. British Machine Vision Conf.,* 2006.

[38] J.J. Zhu, L. Wang, J.Z. Gao, and R.G. Yang, "Spatial-Temporal Fusion for High Accuracy Depth Maps Using Dynamic MRFs," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 5, pp. 899-909, May 2010.
[39] Q. Yang, L. Wang, R. G.Yang, H. Stewenius, and D. Nister, "Stereo Matching with Color-Weighted Correlation, Hierarchical Belief Propagation and Occlusion Handling," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recoginition,* pp. 2347-2354, 2006.
[40] S. Birchfield and C. Tomasi, "A Pixel Dissimilarity Measure That Is Insensitive to Image Sampling," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 4, pp. 401-406, Apr. 1998.
[41] K.J. Yoon and I.S. Kweon, "Locally Adaptive Support-Weight Approach for Visual Correspondence Search," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* pp. 924-931, 2005.
[42] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of Time-of-Flight Depth and Stereo for High Accuracy Depth Maps," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[43] L. Wang, L.H. Jin, and R.G. Yang, "Search Space Reduction for MRF Stereo," *Proc. 10th European Conf. Computer Vision,* 2008.
[44] Y.J. Zheng, S. Lin, and S.B. Kang, "Single-Image Vignetting Correction," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 461-468, 2006.
[45] A. Brunton, C. Shu, and G. Roth, "Belief Propagation on the GPU for Stereo Vision," *Proc. Third Canadian Conf. Computer and Robot Vision,* 2006.
[46] Q.X. Yang, L. Wang, R.G. Yang, S.N. Wang, M. Liao, and D. Nister, "Real-Time Global Stereo Matching Using Hierarchical Belief Propagation," *Proc. British Machine Vision Conf.,* 2006.
[47] C.K. Liang, C.C. Cheng, Y.C. Lai, L.G. Chen, and H.H. Chen, "Hardware-Efficient Belief Propagation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.

Jiejie Zhu received the PhD degree in computer science from State Key Laboratory of CAD and CG, Zhejiang University, China, in 2007. He is currently a postdoctoral researcher in the Computer Science Department, University of Central Florida, and the Perceptural Science Group, Brain and Cognitive Department, Massachusetts Institute of Technology (second coordinator). Earlier, he was a postdoctoral researcher in the Computer Science Department at the University of Kentucky. His research interests include machine learning, computer vision, augmented reality, and human computer interaction. He is a member of the IEEE.



Liang Wang received the BS degree from the School of Computer Science, Beihang University, China, in 2004. He is currently a PhD student in the Computer Science Department at the University of Kentucky. His research interests include computer vision and graphics, especially 3D reconstruction, stereo matching, and image-based modeling. He is a student member of the IEEE.

Ruigang Yang received the MS degree in computer science from Columbia University in 1998 and the PhD degree in computer science from the University of North Carolina, Chapel Hill, in 2003. He is an associate professor in the Computer Science Department at the University of Kentucky. His research interests include computer vision, image processing, and computer graphics, in particular, in the area of 3D modeling and visualization. He is a recipient of the US National Science Foundation CAREER award in 2004, and is a member of the IEEE, the IEEE Computer Society, and the ACM.

James E. Davis received the PhD degree from Stanford University in 2002. He is an associate professor of computer science at the University of California, Santa Cruz (UCSC). His technical research expertise is in ICTD, computer graphics, and machine vision, work that has resulted in more than 80 peer-reviewed publications, patents, and invited talks. He serves as the faculty director of the Center for Entrepreneurship at UCSC, and sits on advisory councils for a handful of startups and nonprofits. Previously, he was a senior research scientist at Honda Research Institute. His has twice received awards for innovative style in his teaching, including a course on the importance of technology to social entrepreneurship. He is a member of the IEEE.

Zhigeng Pan received the bachelor's and master's degrees from the Computer Science Department at Nanjing University in 1987 and 1990, respectively, and the PhD degree in 1993 from Zhejiang University. Since 1993, he has been working at the State Key Laboratory of CAD and CG, Zhejiang University. His research interests include distributed graphics, virtual reality, multimedia, and digital entertainment. He is the editor-in-chief of *The International Journal of Virtual Reality* and co-EiC of *LNCS Transactions on Edutainment*. He is a member of the IEEE.