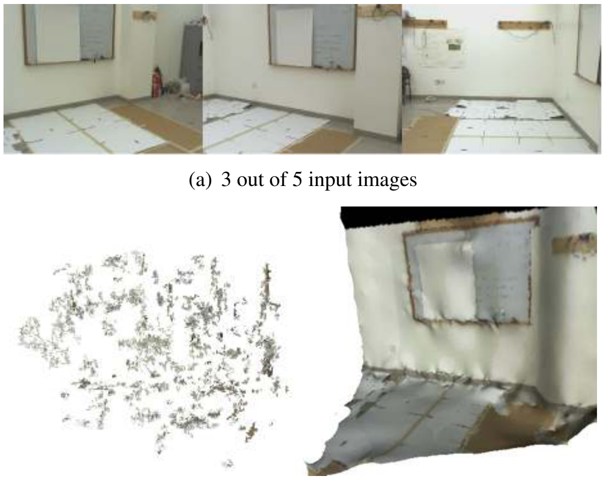
Multi-view Image and ToF Sensor Fusion for Dense 3D Reconstruction

Young Min Kim[1], Christian Theobalt[1], James Diebel[1],
Jana Kosecka[2], Branislav Miscusik[2,3], Sebastian Thrun[1]

Stanford University[1], George Mason University[2], Austrian Institute of Technology GmbH[3]

Abstract

Multi-view stereo methods frequently fail to properly reconstruct 3D scene geometry if visible texture is sparse or the scene exhibits difficult self-occlusions. Time-of-Flight (ToF) depth sensors can provide 3D information regardless of texture but with only limited resolution and accuracy. To find an optimal reconstruction, we propose an integrated multi-view sensor fusion approach that combines information from multiple color cameras and multiple ToF depth sensors. First, multi-view ToF sensor measurements are combined to obtain a coarse but complete model. Then, the initial model is refined by means of a probabilistic multi-view fusion framework, optimizing over an energy function that aggregates ToF depth sensor information with multi-view stereo and silhouette constraints. We obtain high quality dense and detailed 3D models of scenes challenging for stereo alone, while simultaneously reducing complex noise of ToF sensors.



(a) 3 out of 5 input images

(b) Furukawa multi-view stereo (c) our multi-view sensor fusion

Figure 1. *room* data set

1. Introduction

The purely image-based 3D reconstruction of scene geometry, for instance via a stereo method, is still a highly challenging problem. Even the state-of-the-art multi-view stereo methods [7] according to the Middlebury data set evaluation fail to properly reconstruct the 3D scene, Fig. 1(b). The primary reason for this is the notorious difficulty of finding multi-view correspondence when visible texture is sparse or complex occlusions are present. Although these difficulties could be partially remedied by increasing the set of views or resolution of the images, intrinsic problems still remain.

One way to overcome the limitations of image-based reconstruction methods is to combine a conventional multi-view stereo vision system with a flash lidar or Time-of-Flight (ToF) depth sensor [1]. Unlike any other scanner, ToF sensors can capture full frame depth at video frame rates. Therefore they are uniquely suited for capturing 3D information in real time and can greatly advance photo-realistic 3D rendering of dynamic scenes. (Although a sensor of this type is often called a ToF camera to highlight its frame rate, we will call it ToF depth sensor or ToF sensor to avoid confusion with conventional color cameras.) Despite the fact that ToF sensors can provide dense depth maps even where stereo setups typically fail, they have two main challenges: (1) the resolution of ToF depth maps is far below the resolution of stereo depth maps from color images, and (2) measurements are greatly corrupted by non-trivial systematic measurement bias and random noise.

The main contribution of our work is a novel approach for the fusion of multiple ToF sensors with stereo yielding 3D reconstructions superior to the ones obtainable with individual sensing modalities alone. The fusion approach should be able to achieve the accuracy of a stereo approach where possible and the completeness and robustness of a ToF sensor.

At first glance, one might think that a straightforward two-step procedure that first finds initial 3D geometry from ToF sensors, and second applies any state-of-the-art stereo approach produces best results. However, superior results can be achieved by a tighter integration.

We propose an integrated multi-view method that

- utilizes the multi-view setup to compensate complex systematic measurement bias (Sect. 3),
- fuses multi-view ToF measurements into a single complete initial geometry estimate that drastically reduces the random noise by incorporating the directional noise characteristics of ToF sensors (Sect. 5),
- integrates the ToF sensor noise characteristics and stereo cues via probabilistic framework that refines the initial geometry estimate and achieves the accuracy of stereo when reliable constraints exist (Sect. 6).

The framework utilizes both resolution differences and measurement characteristics of the sensors. The level of detail for the initial geometry is limited by the ToF sensor resolution. After refinement step, the reconstruction (Fig. 1(c)) is more detailed and more complete than results obtained with multi-view stereo or single-view fusion.

2. Related Work

Multi-View Stereo Comprehensive surveys of stereo vision techniques can be found in [18, 19]. In general, stereo techniques can be split into two basic categories according to the search range used to calculate the photo-consistency measure. Volumetric carving methods search for a surface in a regular [21] or irregular [15] voxel space. On the other hand, image-based methods estimate the depth for each reference image entity (pixels, lines, windows, or segments) in 3D or along corresponding epipolar lines [19, 7]. Both methods are computationally expensive, and they both result in outliers and holes in areas with repetitive textures, a lack of texture, or substantial lighting changes across the views.

ToF Sensors Unlike other depth sensors, such as laser scanners or structured light scanners, time-of-flight flash lidars [17] or ToF sensors can capture dynamic scenes at real-time frame rates and multiple ToF sensors can run concurrently. Most sensor fusion approaches using ToF sensors aim at enhancing the resolution of depth maps captured with a single sensor, *e.g.* by combining it with a single color camera [5, 22, 14, 4]. This is mainly achieved by enforcing statistical relationships between images and depth, such as collocating intensity and depth discontinuities, and forcing smoothness of geometry in regions of uniform intensity. A few pioneering works of sensor fusion [23] [2] mention fusion of a ToF sensor and a stereo pair of color cameras, but focus only in single-view ToF sensor cases. [8] uses intensity image silhouette and ToF depth data to recover concavities in visual hull reconstructions. Although they extended the use of ToF sensors into a multi-view scenario for 3D reconstruction, their approach does not exploit photo-consistency and yields only marginal improvement over pure visual hull reconstructions at the low resolution level of depth sensors.

In contrast, we propose a new multi-view depth and stereo fusion algorithm that recovers dense multi-view 3D geometry at the high resolution level of color cameras via a Bayesian surface reconstruction technique similar to [6] and [11]. Our system fully utilizes not only the measurement characteristics of ToF depth sensors but also silhouette cues and photo-consistency measures between color cameras.

3. Data Acquisition and Calibration

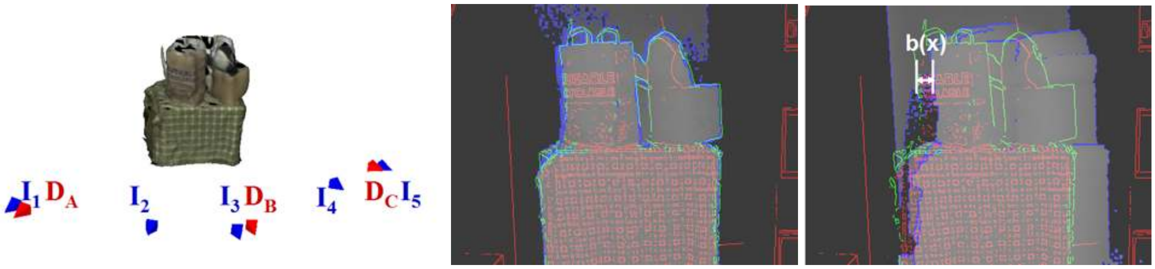
Our multi-view recording system comprises five Point Grey$^{\text{TM}}$ Flea2 color cameras and three MESA Swissranger$^{\text{TM}}$ SR3000 ToF depth sensors (each running at a different modulation frequency to prevent interference). Input to our reconstruction are the data captured from each of the cameras: five intensity images $\mathbf{I} = \{I_1, I_2, \ldots, I_5\}$, each featuring 1024×768 pixels, and the three ToF depth maps $\mathbf{D} = \{D_A, D_B, D_C\}$, each featuring 176×144 pixel depth maps. All cameras and sensors are placed in a semi-circular arrangement around the scene and point roughly towards its center (Fig 2 (a)). Three of the color cameras are paired with three ToF sensors resulting in three pairs with almost identical viewpoints: (I_1, D_A), (I_3, D_B), and (I_5, D_C). As described later, collocating the ToF sensors and color cameras allows us to utilize silhouette constraints.

Since the ToF sensors also provide an intensity image, extrinsic and intrinsic parameters of the ToF and color cameras can be calibrated using a standard calibration technique [3]. However, even if the extrinsic and intrinsic camera parameters were estimated perfectly, the non-trivial error characteristics of the ToF depth sensor will result in misalignments between the depth maps D_A, D_B, and D_C.

The error characteristics consist of two main components: systematic bias and random noise [1, 12, 13]. Following the work of [13], we assume the following measurement model of a ToF sensor:

$$p(z \mid x) \sim \mathcal{N}(z; x + b(x), \sigma^2(x))\,, \qquad (1)$$

where z is the depth measurement along a ray, x is the true distance along that ray, and $b(x)$ is the systematic bias, which in practice mostly depends on the true distance. The systematic bias causes the measurement to consistently deviate from ground truth even if the random noise were canceled out, and therefore, shifts the mean of the random noise distribution. We assume the same bias $b(x)$ regardless of the pixel position within the same sensor, but the value is specific to each depth sensor. The random noise can be



(a)

(b)

(c)

Figure 2. Calibration:(a) Camera set up (b) Alignment of silhouette (green) and the depth edges (blue). The 3D geometry X_B is generated from D_B (blue) and projected on the neighboring view I_3. Silhouette edges (green) are distinguished from ordinary image edges (red) by the projection. (c) Misalignment of the silhouette (green) depth edges (blue) of 3D geometry X_A generated from the oblique sensor D_A demonstrating systematic bias $b(x)$.

faithfully modeled as a normal distribution. The standard deviation $\sigma(x)$ of the random noise is a quadratic function of x as described in [13]. Both the systematic bias and the random noise can be compensated in the multi-view set-up.

The systematic bias component of the error can be calibrated and compensated during pre-processing. Previous work proposed bias calibration from many measurements of a calibration object, *e.g.* a checkerboard [23, 13], which is very tedious in practice. The acquired bias function $b(x)$ is typically composed of both a periodic component and a globally decaying component, and it is hard to model it analytically. In contrast, we propose a new bias calibration and compensation approach that capitalizes on the multi-sensor setup. It only requires a single set of images of an arbitrary scene, and it represents the non-trivial $b(x)$ as a look-up table.

From D_A, D_B, and D_C, we can generate approximate 3D geometry, X_A, X_B, and X_C. By reprojecting the X_i's onto the view of a paired color camera, silhouette edges are extracted from each pair of a depth sensor and a video camera, Fig. 2(b). Among image edges (red), silhouette edges (green) are distinguished by proximity to depth discontinuity of projected X_i (blue). When viewed from an oblique angle, we can estimate the bias that makes the multi-view data consistent, Fig. 2(c). The goal is to find $b(x)$ by getting the true depth edges (blue) to align with the silhouette edges (green).

To serve this purpose, the measured depths $z \in D$ are shifted by discrete distance bias values b_k and the shifted depth map $z - b_k$ is projected into image I from an oblique angle. If a depth pixel projects into the vicinity of silhouette edge in I, the bin counter at the distance $z - b_k$ and the bias b_k is increased. Finally, a look-up table containing one b_k for every observed distance is created by finding the largest matrix entry for each x. This procedure is carried out for each of the ToF sensors separately, yielding three different look-up tables that capture each ToF sensor's specific bias.

After the systematic bias compensation, all the sensors are registered into a common frame, and the depth data align much better in 3D. However, there are still subtle remaining registration inaccuracies and there are still disturbances due to random noise.

4. Overview of Surface Reconstruction

Now that all the sensors are registered into a global coordinate system, our job is to extract the best plausible surface from two different sensors modalities. Even though the reconstruction problem has been studied before for each sensor modality individually, the problem of reconstructing a surface that meets the two kinds of sensor measurements in combination is not well understood.

As mentioned before, the key benefits of ToF sensors are completeness, while multi-view stereo can play an essential role in finer resolution. We propose an integrated method that heavily exploits knowledge of directional sensor noise characteristics which, in the multi-view case, allows for much more reliable 3D localization. In a first initial fusion step, an initial surface estimate $\tilde{X}$ is reconstructed from the multi-view ToF measurements only, Sect. 5. In a refined fusion step, Sect. 6, we exploit specific stereo and image cues to refine the initial fusion result. The image cues we employ are largely motivated by the measurement characteristics of the ToF sensors and can be optimally combined with the sensor uncertainty information for best reconstruction.

5. Initial Surface Reconstruction

3D reconstruction begins by building an initial 3D shape estimate $\tilde{X}$ from the ToF data only. Even after calibration, the raw output of ToF sensors is subject to extreme random noise. In theory, we can reduce the random noise either by (1) taking an average of multiple frames over time or (2) combining ToF data from multiple viewpoints of the same time step while exploiting knowledge about noise characteristics. The former is a trivial and effective method, but is only applicable for static scenes. We take the latter approach that honors the real-time frame rates of ToF sensors and makes our reconstruction method suitable for dynamic scenes, too. As it will be shown later (Sect. 7.2), our approach can reduce the noise level up to the level of the former method.

In the initial surface reconstruction, we merge all depth maps D_A, D_B, and D_C into a common 3D field of occupancy probabilities. We store the combined occupancy values in a regular voxel grid V_d that is aligned with the world coordinate frame and comprises voxels of 1 cm side length (at a scene distance of 2 m, 1 cm voxel size is the spatial resolution of the depth sensor). We describe the occupancy probability for each pixel ray of depth measurement (Sect. 5.1), and a method for fusing the occupancy probabilities from multiple measurement into a joint probability field (Sect. 5.2). From the joint occupancy field, the initial surface is reconstructed via iso-surface extraction. This produces a sub-optimal surface estimate $\tilde{X}$ but is fast and the quality has been found to be sufficient to serve as a starting point for further optimization.

5.1. Occupancy Probabilities

One of the most significant characteristics of ToF sensors is that the measurement uncertainty lies primarily in the direction of each depth ray. For a given measurement z, the most probable distance of the surface along the ray evaluates to $\tilde{x} = \text{argmax}_x \; p(z \mid x) = z - b(x)$, according to the measurement model of Eq.(1). The measurement model of Eq.(1), however, implies directional information, and therefore, cannot be easily applied to integrate multiple sensor measurements from various directions. For example, when a measurement reads $z = 2$ m, the probability according to Eq.(1) represents not only that surface exists at x but also that the one dimensional space along the ray direction in front of the surface is empty, and that we do not know what is behind the surface.

We can use the measurement model to infer a probability of occupancy for each voxel in space independently. $p(m_x \mid z) = 0$ represents the situation that the voxel is completely empty, and $p(m_x \mid z) = 1$ represents a fully occupied voxel. Please note that we purposefully chose to use a heuristic function to transform the measurement model into occupancy probabilities which allows us to build the initial surface more faithfully via rapid iso-surface reconstruction. Our heuristic function incorporates the directional ToF sensor noise characteristics and defines $p(m_x \mid z) = 0.5$ at the most probable distance $\tilde{x}$, Fig. 3(a). The low occupancy probability when $x < \tilde{x}$ represents empty space between the sensor and the measured surface. The steepness of the change reflects the standard deviation of the random noise, or the reliability of the measurement. The probability eventually reaches to 0.5 (occlusion), which is the initial probability without any measurement.

While our heuristic function is similar to the function used in [8], our function is specifically tailored to surface reconstruction via iso-surface extraction. If one uses a classical measurement model as in [8], the occupancy probabilities of the most likely surface vary locally. Therefore, the correct most likely surface can only be found via a costly optimization (e.g. graph-cut) that accounts for these local variations. In contrast, we purposefully modified the measurement model to localize the most probable surface more consistently with an occupancy value of 0.5, also in the multi-view case. We can therefore more faithfully localize an initial surface estimate by performing rapid iso-surface extraction at an iso-level of 0.5. The extracted surface is a starting point for further refinement described in Sect. 6.

5.2. Joint Occupancy and Surface Extraction

The occupancy probability for the voxel grid is initialized to a value of 0.5, which is equivalent to no information. Since we assume that the per-voxel depth measurements z_1, z_2 and z_3 from all three ToF sensors are independent, we can merge them into a joint occupancy probability as follows [20, 9]:

$$\log \frac{p(m_x \mid z_1, z_2, z_3)}{p(\neg m_x \mid z_1, z_2, z_3)} = \sum_{i=1}^{3} \log \frac{p(m_x \mid z_i)}{p(\neg m_x \mid z_i)} \qquad (2)$$

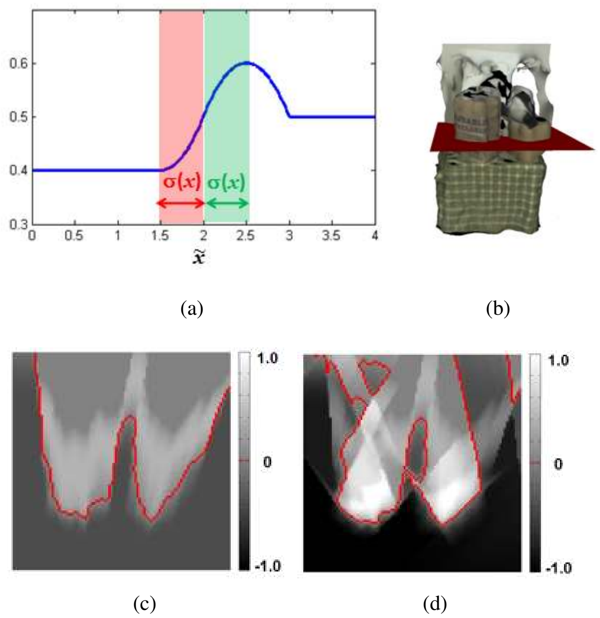
Here $p(\neg m_x \mid z_i) = 1 - p(m_x \mid z_i)$ and log-odds are used for numerical accuracy. Starting from uniform occupancy probability of $p(m_x \mid z) = 0.5$, the directional uncertainty around the iso-value is reduced as three sensor measurements are combined, Fig. 3 (c)(d). As a consequence, the surface extracted via Marching Cubes iso-surface extraction is much closer to the true surface than if we had used one ToF sensor only.

6. Detailed Surface Reconstruction

Now that we have a plausible initial surface estimate, we can use a gradient-based optimizer to refine the surface position. The reconstruction of a 3D geometry model X is formulated as the problem of finding the most likely (MAP) surface given the ToF depth measurements Z, 3D point constraints according to multi-view photo-consistency C, and 3D positions constraints S due to occlusion boundaries that should line up with image discontinuities. Assuming the independence of the measurement likelihoods of ToF data, stereo constraints and silhouette constraints, we can formulate the posterior probability of the 3D model given the three types of measurements as:

$$\mathbf{P}(X \mid Z, C, S) \propto \mathbf{P}(Z \mid X)\mathbf{P}(C \mid X)\mathbf{P}(S \mid X)\mathbf{P}(X)\,. \qquad (3)$$

Here, $\mathbf{P}(Z \mid X)$, $\mathbf{P}(C \mid X)$, and $\mathbf{P}(S \mid X)$ are the measurement likelihoods of Z, C and S, and $\mathbf{P}(X)$ is a prior



(a) (b)

(c) (d)

Figure 3. (a) Heuristic occupancy probability function of each ray. For a 3D model, a slice of occupancy probability field at (b) is shown for (c) one sensor and (d) three sensors in log odd scale. The corresponding initial surface where $p(m_x \mid z) = 0.5$ is shown in red line. (In log odd scale, $p(m_x \mid z) = 0.5$ corresponds to $\log \frac{p(m_x|z)}{p(\neg m_x|z)} = 0$.) Note the improved surface localization when combining of multiple ToF sensors.

on likely 3D model configurations. The MAP estimate of 3D model $\hat{X}$ is found by minimizing the negative logarithm of the above posterior which yields an energy minimization problem of the form:

$$\hat{X} = \underset{X}{\operatorname{argmin}}\, E_Z + E_C + E_S + E_X \qquad (4)$$

where E_Z, E_C, and E_S are the negative log-likelihoods of $\mathbf{P}(Z \mid X)$, $\mathbf{P}(C \mid X)$, and $\mathbf{P}(S \mid X)$, respectively. Accordingly $E_X = -\log \mathbf{P}(X)$. We represent the most likely 3D model X as a triangle mesh with fixed vertex connectivity, and thus $X = \{\mathbf{v}_i \mid i = 1, \ldots, N\}$ can be interpreted as the set of all N vertex positions $\mathbf{v}_i$ of the mesh. In the following, we explain the components of the energy function in more detail.

6.1. Measurement Potential for ToF Sensors

The measurement potential E_Z tries to hold the vertex positions of $\hat{X}$ close to $\tilde{X}$, but also takes into account the resolution deficiency of the initial estimate by assigning no penalty to any surface that only moves within one voxel size $\delta = 1$cm from $\tilde{X}$.

$$E_Z = \sum_{i=1}^{N} \|\mathbf{r}_i\|^2, \qquad (5)$$

where $\mathbf{r}_i = (r_i^x, r_i^y, r_i^z)^\top$ contains the distances along x,y and z between the current vertex position $\mathbf{v}_i \in X$ and the position of the same vertex according to the initial reconstruction $\tilde{\mathbf{v}}_i \in \tilde{X}$. Using r_i^x as an example, the distances evaluate to $r_i^x = \max(0, |v_i^x - \tilde{v}_i{}^x| - \delta/2)$, with r_i^y and r_i^z computed accordingly. The optimizer can thus freely move the surface within the 3D space of one voxel size. This way we successfully recover finer surface detail from the intensity camera data. Simultaneously, we remove surface discretization artifacts from the initial surface estimates that are due to limited sensor and thus voxel grid resolution.

6.2. Photo-consistency and Silhouette Potentials

When reliable stereo cues exist, we can refine our reconstruction. Among many possible choices, photo-consistency measure and silhouette constraints are used for high-resolution constraints. By enforcing the constraints locally, we can overcome the subtle remainders of sensor noise, small inaccuracies in initial surface reconstruction, and limited depth camera resolution. Note that we were able to enforce silhouette constraints without a complex segmentation algorithm because we already have a good approximation of 3D geometry from ToF sensor measurements.

To extract 3D points to be used as photo-consistency constraint, we first test reliability for every vertex $\tilde{\mathbf{v}}_i \in \tilde{X}$ by checking if there is sufficient color variation with respect to the most fronto-parallel camera. If there is, the point $\tilde{\mathbf{v}}_i$ is transformed along the normal direction within the vicinity of the initial reconstruction and the photo-consistency is calculated over a patch centered at the transformed point $\mathbf{p}_k$. If the photo-consistency of the transformed point is a local maximum larger than a threshold value (0.5), the pair $(\mathbf{v}_i, \mathbf{p}_k)$ is added to the list C of stereo-based constraints. Given the list of all such constraints, we formulate the following photo-consistency potential:

$$E_C = \sum_{(\mathbf{v}_i, \mathbf{p}_k) \in C} \mathcal{H}(\|\mathbf{v}_i - \mathbf{p}_k\|)\,. \qquad (6)$$

Here, $\mathcal{H}(x)$ is the robust Huber regression function [10]. The transition position where the Huber function switches to an ℓ_1 norm is at $x = \delta/2$. This way, we implicitly downweight the influence of obvious outliers on the final reconstruction because it is unlikely that the surface ought to deform by more than one voxel size.

Similarly, we add a silhouette potential E_S to drive vertices which lie on a geometric occlusion boundary to the nearest reprojected intensity boundary. For each vertex $\tilde{\mathbf{v}}_i \in \tilde{X}$, we check if it is an occlusion boundary with respect to each color camera. If it is, $\tilde{\mathbf{v}}_i$ is displaced along its local normal direction in a neighborhood around its original position, yielding new 3D candidate positions $\mathbf{p}_u$. If a transformed position $\mathbf{p}_u$ is found that projects into an edge in a color image, the pair $(\mathbf{v}_i, \mathbf{p}_u)$ is added to the list S of silhouette rim constraints. Given the set S, an identical expression as described in Eq.(6) can be used to formulate E_S.

6.3. Prior Potential

The prior potential E_X serves as a regularizer that favors likely 3D surface configurations. We resort to a Laplacian prior [7] that reads as follows:

$$E_X = \sum_{i=1}^{N} \|\rho \Delta \mathbf{v}_i + (1-\rho)(-\Delta^2 \mathbf{v}_i)\|^2 \,. \tag{7}$$

Here, $\Delta \mathbf{v}_i$ is the discrete Laplace operator evaluated at vertex $\mathbf{v}_i$, and $\Delta^2 \mathbf{v} = \Delta(\Delta \mathbf{v}_i)$ is the respective bi-Laplacian. Through experiments we could verify that $\rho = 0.6$ produces the best results.

7. Results

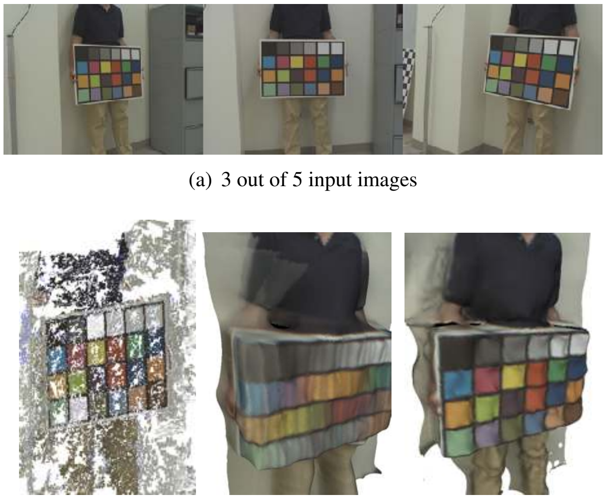
We show results with four different data sets recorded with the setup described in Sect. 3. Henceforth we refer to the data sets as *room* (Fig. 1), *macbeth* (Fig. 4), *girl* (Fig. 5), and *whale* (Figs. 6).

We use the L-BFGS-B optimizer to solve the final energy minimization problem [16]. Currently, we employ a single-threaded non-optimized C++ implementation that takes around 15 minutes to create a final reconstruction on a Dual Core Athlon$^{\text{TM}}$ 5600+ machine with 4 GB memory, excluding the bias calibration step. Run times are dominated by the photo-consistency calculation (11-12 min) which can be drastically sped up, e.g. by using the GPU. Overall, we expect that run times in the range of 4-5 minutes per frame are feasible through code optimization.

7.1. Conceptual Advantage over Multi-view Stereo

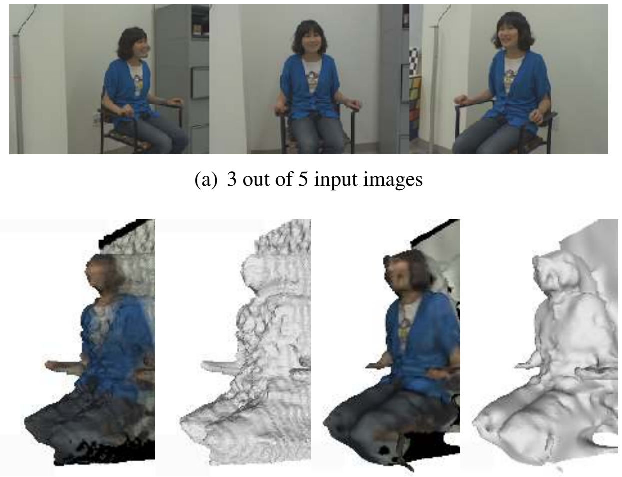
Even the state-of-the-art multi-view stereo approach [7] fails on our *room* data set (Fig. 1(b)) which features mainly plain untextured walls. In contrast, our multi-view sensor fusion method can capitalize on and refine the geometry measured with the ToF sensors, thereby reconstructing dense and faithful models of the back walls and the floor, Fig. 1(c).

Our other data sets also feature elements that are difficult for a multi-view stereo methods. For instance, in the *whale* data set (Fig. 6(b) left), the whale itself, the box, and the bag exhibit big holes, since the surfaces are uniformly colored. The same holds true for the colored board and the t-shirt in *macbeth* (Fig. 4(b)). The holes of multi-view stereo reconstruction can be filled based on a smoothness prior (e.g., Poisson surface reconstruction). However, geometric hole-fillling approaches are not based on true measurement, and can hallucinate incorrect geometry in between objects (Fig. 6(b) right). In contrast, our proposed algorithm yields accurate dense models even on such challenging scenes, Figs. 4(d), 6(c), 5(c).


(a) 3 out of 5 input images

(b) Furukawa multi-view stereo (c) one ToF sensor and a stereo pair (d) our multi-view sensor fusion

Figure 4. *macbeth* data set

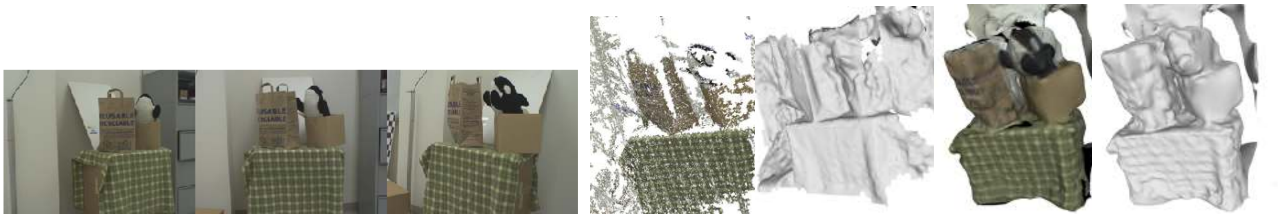

(a) 3 out of 5 input images

(b) one ToF sensor and a stereo pair (c) our multi-view sensor fusion

Figure 5. *girl* data set

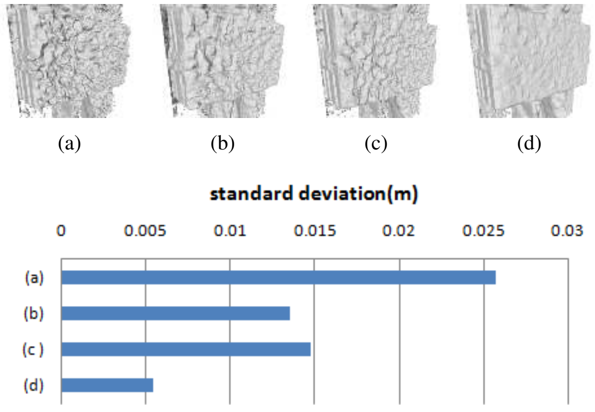
7.2. Conceptual Advantage over Single-View Sensor Fusion

Our proposed fusion of information from multiple ToF sensors and multiple intensity cameras yields results of superior quality than it is achievable with methods employing only a single ToF sensor and a single pair of vision cameras [23, 2]. First, our approach captures a larger range of viewpoints yielding more complete geometry with less occlusion problems. Furthermore, our multi-view approach enables more accurate 3D geometry by efficiently compensating for both systematic bias and random noise. To illus-



(a) 3 out of 5 input images (b) Furukawa multi-view stereo (c) Multi-view fusion

Figure 6. *whale* data set



(e)

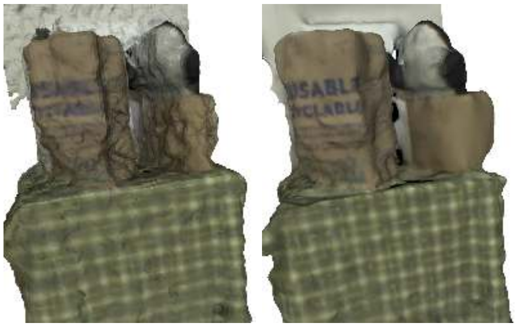
Figure 7. Random noise levels on a planar surface which is part of the *macbeth* data set: The very high random noise label in a depth map from a single ToF sensor (a) is severely reduced by our multi-view approach (b). Please note that by combining three frames from three ToF sensors taken at the same instant (c), we achieve similar noise reduction as if we temporally averaged three consecutive frames from a single camera. This underpins the feasibility of our multi-view approach also for dynamic scenes. For comparison, the result of temporally averaging 100 frames from one sensor is shown in (d). The quantitative result is shown in (e) which plots the standard deviations of random noise for the respective methods.

trate this, we compared our approach to a single-view approach by running sensor fusion with one ToF sensor and two adjacent video cameras only, as shown in Fig. 4(b) and Fig. 5(b). The single-view results exhibit erroneous extruded occlusion boundaries whereas our algorithm reconstructs geometry faithfully also in those areas occluded from one viewpoint (such as the fins of the whale, or the arms of the girl). Furthermore, when reprojecting the input intensity images back onto the 3D geometry, there are strongly noticeable ghosting artifacts in the single view results which are not visible in our results; an indication for the more accurate 3D reconstruction through our approach.

Fig. 7 shows that our multi-view formulation in Sect. 5 can reduce the severe random noise from a single ToF sensor.

data set	NCC		silhouette dist.	
	initial	refined	initial	refined
whale	0.3320	0.6219	4.333	1.377
macbeth	0.5694	0.7047	3.147	1.110
girl	0.4703	0.5904	4.349	1.323
room	0.6385	0.8020	1.581	0.5799

Table 1. On all data sets, the refinement step leads to a clear improvement of the average photo-consistency (NCC) of the vertices which were included into the refinement according to photo-consistency constraints. The silhouette in images and depth discontinuity of 3D reconstruction is also closer after refinement. This quantitatively confirms the observable visual improvements after refinement of the coarse initial geometry.

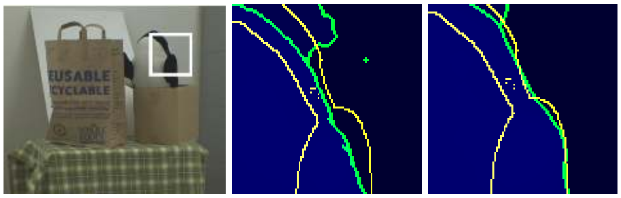


(a) (b)

Figure 8. Improvement in Texture (Blended from Input Images): The initial model (a) shows reconstruction errors and texture ghosting, which is significantly improved in the full fusion result (b). The patterns in the table cloth is more clear and the text is readable.

7.3. Verification of Pipeline and Quantitative Validation

Our multi-view pipeline extracts faithful and complete 3D models, even on sparsely textured scenes. This is achieved through an efficient interplay of initial reconstruction and refinement under exploitation of ToF sensor characteristics. The relevance of the refinement step for accurate 3D geometry computation is revealed after visual inspection on texture, Fig. 8, and silhouette alignments, Fig. 9). This clearly illustrates the effectiveness of multi-view ToF-guided enforcement of photo-consistency and sil-



(a) (b) (c)

Figure 9. Alignment Improvement: Zooming in onto the whale (white box) one can see that, after the full pipeline (c), the whale's occlusion edges in 3D (green) align much more accurately with the corresponding image edges (yellow) than after initial reconstruction using ToF sensor only (b).

houette constraints for accurate shape reconstruction.

Table 1 shows that the full pipleline clearly improves reconstruction quality over mere initial ToF-based reconstruction on all data sets. Please note that, since we did not have a 3D laser scanner for ground truth geometry measurement, we resorted to NCC and silhouette distances for quantitative validation.

8. Conclusion

We have presented a new multi-view sensor fusion algorithm that combines multiple ToF depth measurements and multiple color images of a scene to reconstruct accurate and dense 3D models. Our fusion framework is designed in the way that the mutual benefits of both sensor types can be most pronounced. The final reconstruction produce a faithful models even of scenes where multi-view stereo or single-view sensor fusion fails. Since multiple ToF sensors can run together at full video frame rate (in contrast to other depth sensors), we plan to apply our technique to time-varying data in the future.

References

[1] D. Anderson, H. Herman, and A. Kelly. Experimental characterization of commercial flash ladar devices. In *International Conference of Sensing and Technology*, 2005.

[2] C. Beder, B. Bartczak, and R. Koch. A combined approach for estimating patchlets from PMD depth images and stereo intensity images. In *DAGM*, pages 11–20, 2007.

[3] J. Bouguet. http://www.vision.caltech.edu/bouguetj/calib_doc/, 2007.

[4] D. Chan, H. Buisman, C. Theobalt, and S. Thrun. A Noise-Aware Filter for Real-Time Depth Upsampling. In *Proc. of ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, pages 1–12, 2008.

[5] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In *NIPS*, pages 291–298. 2006.

[6] J. Diebel, S. Thrun, and M. Brünig. A bayesian method for probable surface reconstruction and decimation. *ACM Transactions on Graphics*, 25:39–59, January 2006.

[7] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *Proc. of CVPR*, June 2007.

[8] L. Guan, J.-S. Franco, and M. Pollefeys. 3d object reconstruction with heterogeneous sensor data. *3DPVT08*, 2008.

[9] C. Hernandez, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. In *CVPR*, 2007.

[10] P. Huber. *Robust Statistics*. Wiley, 2004.

[11] P. Jenke, M. Wand, M. Bokeloh, A. Schilling, and W. Straer. Bayesian point cloud reconstruction. *Computer Graphics Forum*, 25(3):379–388, 2006.

[12] T. Kahlmann, F. Remondino, and H. Ingensand. Calibration for increased accuracy of the range imaging camera swissrangertm. In *Proc. of IEVM*, 2006.

[13] Y. M. Kim, D. Chan, C. Theobalt, and S. Thrun. Design and calibration of a multi-view TOF sensor fusion system. In *Proc. of CVPR Workshop on Time-of-flight Computer Vision*, pages 1–7, 2008.

[14] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele. Joint bilateral upsampling. *ACM TOG*, 26(3), 2007.

[15] P. Labatut, J. P. Pons, and R. Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *ICCV*, 2007.

[16] J. Nocedal. Updating quasi-newton matrices with limited storage. In *Mathematics of Computation*, volume 25, pages 773–782, 1980.

[17] T. Oggier, M. Lehmann, R. Kaufmann, M. Schweiz er, M. Richter, P. Metzler, G. Lang, F. Luste nberger, and N. Blanc. An all-solid-state optical range camera for 3d real-time imaging with sub-centimeter depth resolution. In *Proc. SPIE: Optical Design and Engineering*, pages 534–545, 2004.

[18] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47:7–42, 2002.

[19] S. Seitz, B. Burless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. of CVPR*, volume 1, pages 519–528, 2006.

[20] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[21] G. Vogiatzis, C. H. Esteban, P. H. S. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts and occlusion robust photo-consistency. *PAMI*, 29(12):2241–2246, 2007.

[22] Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *CVPR*, 2007.

[23] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *Proc. of CVPR*, 2008.