3D Object Reconstruction with Heterogeneous Sensor Data

Li Guan, Jean-Sébastien Franco, Marc Pollefeys

3D Object Reconstruction with Heterogeneous Sensor Data

Li Guan
UNC-Chapel Hill, USA
ETH-Zurich, Switzerland
lguan@cs.unc.edu

Jean-Sebastien Franco
LaBRI - INRIA Sud-Ouest
University of Bordeaux, France
jean-sebastien.franco@labri.fr

Marc Pollefeys
UNC-Chapel Hill, USA
ETH-Zurich, Switzerland
marc@cs.unc.edu

Abstract

In this paper, we reconstruct 3D objects with a heterogeneous sensor network of Time of Flight (ToF) Range Imaging (RIM) sensors and high-res camcorders. With this setup, we first carry out a simple but effective depth calibration for the RIM cameras. We then combine the camcorder silhouette cues and RIM camera depth information, for the reconstruction. Our main contribution is the proposal of a sensor fusion framework so that the computation is general, simple and scalable. Although we only discuss the fusion of conventional cameras and RIM cameras in this paper, the proposed framework can be applied to any vision sensors. This framework uses a space occupancy grid as a probabilistic 3D representation of scene contents. After defining sensing models for each type of sensors, the reconstruction simply is a Bayesian inference problem, and can be solved robustly. The experiments show that the quality of the reconstruction is substantially improved from the noisy depth sensor measurement.

1. Introduction

3D object reconstruction is a classic computer vision problem and has many applications such as virtual reality, vision-guided surgeries, medical studies and simulations, video games, architectural design, etc. Within the past five years, a promising new technology, Range Imaging (RIM) cameras based on Time of Flight (ToF) principles are coming to market. Swiss Ranger 3000 as shown in Fig. 1 is a typical model. 2.5D range images combined with 2D intensity images can be directly read out up to 50 fps. Although most of these RIM cameras do not have high image resolution (e.g. 176 x 144 for Swiss Ranger 3000), their measurement throughput is still far beyond the traditional depth sensors, such as LIDAR. This opens enormous potential in a wide range of application areas, including action recognition and tracking, object pose recognition, obstacle detection and so on. However, few literatures have explored its potential in 3D object reconstruction. The main challenges are (1) the range images are noisy and not always accurate enough for 3D reconstruction purposes. In fact, the RIM camera depth calibration itself remains a new and active research topic [15]; (2) the relatively low image resolution prohibits detailed reconstruction.

Figure 1. (a) RIM camera, Swiss Range 3000 (b) a typical output from the sensor.

In this paper, we propose to solve the above problems and explore the reconstruction potential of the RIM cameras by introducing a heterogeneous sensor network of RIM cameras and high-res camcorder. We first describe a simple but effective depth calibration process for the RIM cameras, which is just an additional step in the classic geometric camera calibration procedure [28]. For the reconstruction, we propose a probabilistic framework to fuse depth information from RIM cameras and silhouette cues from camcorders.

The depth information and silhouette cues alone have been explored intensively for 3D reconstruction purpose. Both have their own advantages and drawbacks. For the depth information, it can give you actual object surface patches. But due to self occlusion, individual patches only provide a partial model of the object surface, so one of the many challenges is to deal with missing patches and fill up the holes so as to get a topologically correct object shape [1, 6, 13, 7, 4]. On the other hand, reconstruction from silhouette cues [2, 16, 25, 10, 17] are praised for a closed-form shape estimate of the object. And recently even no hard-decision binary silhouette images are required for a robust

probabilistic visual hull reconstruction [11]. An inherent drawback of a visual hull is that it cannot recover object concavities no matter how many views of silhouettes are provided. However, this can be directly compensated by the depth information. In fact, object depth and silhouette are quite complementary information in nature: the former encodes lights bouncing back from the frontal surfaces; and the latter is tangent to the object. So in theory, these two could be combined to improve the reconstruction quality. Additionally, in our sensor network, the shape details can be recovered with the high-res camcorder frames to compensate the low-res RIM camera images.

However, silhouette and depth integration is not straightforward due to the heterogeneity of the information. Li *et.al.* try to tackle the problem with pure surface representation [18], which requires a lot of delicate handling of geometry computation errors. As an alternative, volumetric fusion can be favored to avoid topological problems [20, 23, 27], but these methods are all based on deterministic criterions, which have to specifically deal with sensor noise perturbations.

In order to achieve a more robust but also more general solution to the fusion problem, similar to [11], our framework borrows the concept of a space occupancy grid from the robotics literature [8, 21, 22] as the representation of 3D scenes. After defining the probabilistic sensing models for each type of sensors, the reconstruction simply becomes a Bayesian inference. The reconstruction result is a posterior probability volume given sensor observations. It is inherently robust and requires no special treatment regarding sensor noise, because the noise and variation is already part of the probabilistic sensing models. One thing to note is that the proposed framework is not limited to the fusion between silhouette cues and depth maps, but any type of sensor observations such as point clouds and disparity maps, as long as the sensing model can be properly defined.

The paper is organized as follows: In Section 2 we explain the mechanism for common RIM cameras and introduce our calibration method. Then in Section 3, we formally describe our reconstruction algorithm via the Bayesian inference framework. And we introduce the camcorder and RIM camera sensing models in Section 4. In Section 5 we validate the proposed calibration and fusion scheme by reconstructions from two real-world datasets.

2. RIM Camera and its Calibration

Most common RIM camera designs, including the SR 3000 in Fig. 1, are based on the indirect ToF-principle. Amplitude modulated light is emitted from the camera, travels to the object, is reflected, and finally demodulated by means of a specialized CMOS/CCD pixel. Demodulation is known as the reconstruction of the received signal, as shown in the

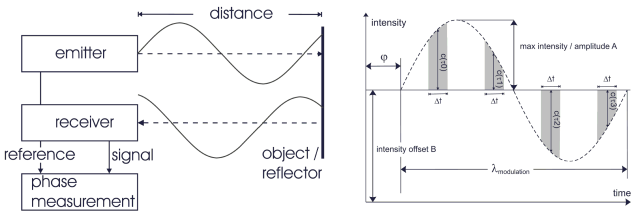

Figure 2. Left: indirect ToF principle. Right: phase shift computation. Adapted from [15]

Fig. 2. In case of a sinusoidal signal, three parameters have to be calculated: the intensity B, the amplitude A, and the phase ϕ. Four sampling points τ_0, τ_1, τ_2 and τ_3 (intensity measurements, each shifted of $90°$) are triggered to the emitted wave. Thus the phase shift ϕ can be calculated as:

$$\phi = \arctan(\frac{\tau_1 - \tau_3}{\tau_0 - \tau_2}) \tag{1}$$

The phase shift ϕ is directly proportional to the distance D the light has traveled:

$$D = \frac{\lambda_{mod}}{2} \cdot \frac{\phi}{2\pi} \qquad \text{with } D < \lambda_{mod} \tag{2}$$

The main issue with the depth measurement is that the measures have low accuracy (the displacement between the ground truth and the measured depth is large). For example, the SR 3000 may have a $0.35m$ displacement aiming at an object at $5.5m$. For 3D reconstruction, this displacement is not negligible and needs to compensated in advance. It can be done by the depth calibration, but so far, only very delicate and expensive devices [15, 19] are available to do so.

Because (1) it is not feasible to have such expensive and sophisticated calibration devices and (2) the calibration is beyond our 3D reconstruction accuracy requirement, we hereby propose an easier calibration procedure taking advantage of our heterogeneous camera network design. Firstly, since the RIM cameras can also produce intensity images, all the cameras can be geometrically calibrated using Bouguets toolbox based on [28]. After the bundle adjustment, we know the relative pose of the cameras and the calibration checkerboard patterns in 3D space with absolute scale. The error is normally smaller than $1cm$, which is reasonable for our 3D reconstruction of human-size figures. Then we perform the depth calibration of the RIM camera simply by finding out a mapping function between these checkerboard pattern poses as the ground truth and the depth readouts from the RIM camera. A detailed example is given in Section 5.

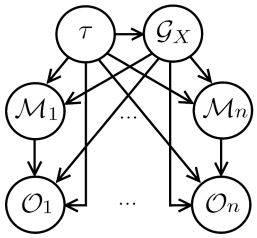

Figure 3. General system dependency.

3. Problem Formulation

Assume we have a set of calibrated sensors, in this section we introduce our probabilistic shape inference framework in details. With the following notations, we can define our problem formally: given a set of synchronized observations $\mathcal{O}$ from n sensors at a specific time instant, we infer for every 3D location X in an occupancy grid $\mathcal{G}$ expanding the 3D space its probability of being occupied or not by the object that we are modeling. And we denote this probability as $p(\mathcal{G}_X)$ with $\mathcal{G}_X$ the binary variable at X.

Notations

n	total number of sensors
X	3D location
i	sensor index
p	pixel index in sensor i corresponding to X
τ	prior
$\mathcal{G}_X$	voxel occupancy at X in the occupancy grid $\mathcal{G}$
$\mathcal{O}_i^p$	observation at pixel p of sensor i
$\mathcal{M}_i$	sensor i's model, in this paper we specifically consider consumer camcorders and 3D depth camera
$\mathcal{B}$	camcorder's background model
$\mathcal{D}$	RIM camera's depth observation model
$\mathcal{S}_i^p$	silhouette formation at pixel p of camcorder i
$\mathcal{T}_i^p$	object's front most surface location with respect to pixel p of RIM sensor i
$\mathcal{P}$	silhouette sampling variable
$\mathcal{R}$	silhouette detection external cause

An intuitive assumption made throughout this paper is that space occupancy variable $\mathcal{G}_X \in \{0, 1\}$ depends only on the information along optic rays that go through X. However, anti-aliasing effects need to be considered. We simply use the same sampling window strategy introduced in [11], where a certain 3D voxel affects the formation of pixels within the sampling window similar to a point spread function. Another common occupancy grid assumption as in [8] is that, we assume statistical independence between voxel occupancies, and compute each voxel occupancy likelihood independently for tractability. Results show that independent estimation, while not as exhaustive as a global search over all voxel configurations, still provides very robust and usable information, at a much lower cost. Therefore, we model the sensor network relationships as computing the joint probability of these variables, $p(\mathcal{G}_X, \mathcal{O}_{1,\ldots,n}, \mathcal{M}_{1,\ldots,n}, \tau)$, and propose the following decomposition, based on the statistical dependencies expressed in Fig. 3:

$$\begin{aligned} &p(\mathcal{G}_X, \mathcal{O}_{1,\ldots,n}, \mathcal{M}_{1,\ldots,n}, \tau) \\ =&p(\tau)p(\mathcal{G}_X|\tau)\prod_{i=1}^{n} p(\mathcal{M}_i|\tau)p(\mathcal{O}_i|\mathcal{G}_X, \mathcal{M}_i, \tau) \end{aligned} \tag{3}$$

- $p(\tau)$ represents the prior probabilities of our parameter set. Since we have no a priori reason to favor any parameter values, we set it to a uniform distribution. It thus disappears from any subsequent inference.
- $p(\mathcal{G}_X|\tau)$ is the prior likelihood for occupancy, which is independent of all other variables except τ. We choose not to favor any voxel location and set this term to uniform in this paper.
- $p(\mathcal{O}_i|\mathcal{G}_X, \mathcal{M}_i, \tau)$, or more specifically, given our aforementioned viewing-ray independence assumption, $p(\mathcal{O}_i^p|\mathcal{G}_X, \mathcal{M}_i^p, \tau)$ represents the sensor observation probability.

Once the joint probability distribution has been fully determined, it is possible to use Bayes rule to infer the probability distributions of our searched variable $\mathcal{G}_X$, given the sensor models $\mathcal{M}$ and their observations $\mathcal{O}$.

$$\begin{aligned} &p(\mathcal{G}_X|\mathcal{O}_{1,\ldots,n}, \mathcal{M}_{1,\ldots,n}, \tau) \\ =&\frac{\prod_{i=1}^{n} p(\mathcal{O}_i^p|\mathcal{G}_X, \mathcal{M}_i^p, \tau)}{\sum_{\mathcal{G}_X}\prod_{i=1}^{n} p(\mathcal{O}_i^p|\mathcal{G}_X, \mathcal{M}_i^p, \tau)} \end{aligned} \tag{4}$$

If we apply Eq. 4 for all locations and obtain this probabilistic volume $\mathcal{G}$, we can simply reconstruct our 3D objects by extracting iso-probability surfaces, or more robustly using state-of-the-art techniques, such as Graphcut/Levelset algorithms [24, 26]. The remaining problem is to define the proper sensor models $\mathcal{M}$ so that the observation formation $p(\mathcal{O}_i|\mathcal{G}_X, \mathcal{M}_i, \tau)$ in Eq. 4 is reasonable. But so far, we have introduced a very general sensor fusion framework, which has no constraints on the sensor type nor data type.

4. Sensor Models

In this section, we describe the probabilistic camcorder background model $\mathcal{B}$ and RIM camera depth model $\mathcal{D}$, which are used in our sensor network. Namely, we analyze the components of $p(\mathcal{O}_i^p|\mathcal{G}_X, \mathcal{B}_i^p, \tau)$ and $p(\mathcal{O}_i^p|\mathcal{G}_X, \mathcal{D}_i^p, \tau)$ for the two types of sensors respectively.

4.1. Camcorder Sensor Model

The sensor observation $\mathcal{O}_i^p$ for a camcorder is the color or intensity but not the silhouette $\mathcal{S}_i^p \in \{0, 1\}$ of the object being reconstructed. However, they are directly related as shown in Fig. 4: the existence of an object at $\mathcal{G}_X$ determines the value of the object silhouette $\mathcal{S}_i^p$. The state of $\mathcal{S}_i^p$ as well as the background color appearance $\mathcal{B}_i^p$ determine the color to be observed.

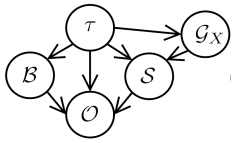

Figure 4. Camcorder dependency. Adapted from [11].

Since $\mathcal{S}_i^p$ is not the direct observation from the camera, it is a latent variable and is marginalized as shown in Eq.5.

$$\begin{aligned} &p(\mathcal{O}_i^p|\mathcal{G}_X, \mathcal{M}_i^p, \tau) \\ &= \sum_{\mathcal{S}_i^p} p(\mathcal{O}_i^p|\mathcal{S}_i^p, \mathcal{B}_i^p, \tau) p(\mathcal{S}_i^p|\mathcal{G}_X, \tau) \end{aligned} \tag{5}$$

- $p(\mathcal{O}_i^p|\mathcal{S}_i^p, \mathcal{B}_i^p, \tau)$ is the **image formation term**. If $\mathcal{S}_i^p = 0$, then $\mathcal{O}_i^p$ can be explained by the background model $\mathcal{B}_i^p$. In this paper, we model it as an RGB color space normal distribution $\mathcal{N}(\mathcal{O}_i^p, \mu_i^p, \sigma_i^p)$, where (μ_i^p, σ_i^p) are the distribution parameters, and trained in advance from a number of images with only empty scene but no reconstruction object in the presence. If $\mathcal{S}_i^p = 1$, the pixel should display the foreground object's color. We set it to a uniform distribution $\mathcal{U}_i^p$, meaning any color can be possibly observed from the object that we are reconstructing. This sensor model is consistent with [11], which follows the basic *background subtraction algorithm* [12, 9], without forcing a hard decision of a binary silhouette image, thus to be much more robust against sensor noise and environment lighting variations.

- $p(\mathcal{S}_i^p|\mathcal{G}_X, \tau)$ is the **silhouette formation term**. It models the silhouette detection response of a single pixel sensor (i, p) to the occupancy state of $\mathcal{G}_X$. In our discretized world, the assumption that a voxel lies on the viewing line of a pixel is uncertain. This may be due to many external causes: potential camera calibration errors, camera mis-synchronization etc. This can be modeled by a latent variable — the sampling variable $\mathcal{P}$. Second, there can be causes for silhouette detection other than the voxel itself: an object occupancy other than the one related by $\mathcal{G}_X$, which is modeled by another hidden variable — external detection cause $\mathcal{R}$. The complete dependencies are shown in Fig. 5. This relationship model is introduced by [11], where detailed formulations regarding these variables can be found.

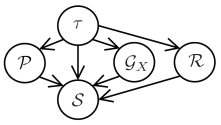

Figure 5. Silhouette detection dependency. Adapted from [11].

4.2. RIM Camera Sensor Model

For a RIM camera, the observation $\mathcal{O}_i^p$ is the depth measurement. Similar to the silhouette variable $\mathcal{S}_i^p$ in camcorder sensor, here we also introduce a latent variable $\mathcal{T}_i^p$, to model the front most surface of the object with respect to the RIM camera. The relationship between sensor variables is shown in Fig. 6. Basically, the existence of an object at $\mathcal{G}_X$ affects the front most surface location $\mathcal{T}_i^p$ to a certain RIM camera i. And $\mathcal{T}_i^p$ affects the depth measurement directly.

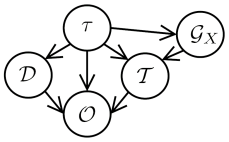

Figure 6. RIM camera dependency.

Because $\mathcal{T}_i^p$ is a latent variable, we also need to marginalize it. However, $\mathcal{T}_i^p$ is not a binary variable as its counterpart — the silhouette $\mathcal{S}_i^p$ for a camcorder, but its range expands all possible locations along the viewing direction. Namely, $\mathcal{T}_i^p \in [0, d_{max}]$, with 0 being the RIM camera optical center, and d_{max} the largest detectable distance of the RIM camera.

$$\begin{aligned} &p(\mathcal{O}_i^p|\mathcal{G}_X, \mathcal{M}_i, \tau) \\ &= \int_0^{d_{max}} p(\mathcal{O}_i^p|\mathcal{T}_i^p, \mathcal{D}_i^p, \tau) p(\mathcal{T}_i^p|\mathcal{G}_X, \tau) d\mathcal{T}_i^p \end{aligned} \tag{6}$$

- $p(\mathcal{O}_i^p|\mathcal{T}_i^p, \mathcal{D}_i^p, \tau)$ is the **depth measurement term**. It depicts how precise the RIM camera depth measure is. We use a normal distribution $\mathcal{N}(\mathcal{T}_i^p, \sigma)$ to model it, where σ is trained from depth calibration process or obtained from the camera manual.

- $p(\mathcal{T}_i^p|\mathcal{G}_X, \tau)$ is the **surface formation term**. Assume every voxel is independent along the viewing direction

of length d_{max}, and any place on the viewing ray has an equal chance of $1/d_{max}$ to be the front most point. Now, if $\mathcal{G}_X = 1$, the front most surface position $\mathcal{T}_i^p$ still has a chance of $1/d_{max}$ to be at any position in front of X, namely $\mathcal{T}_i^p < d_X - \epsilon$, where $\epsilon \to 0$. But this is not the case for the positions behind X, because X is already blocking the viewing ray. Eq. 7 & 8 shows the complete scenario, with d_X being the distance from X to the RIM camera. Both distributions of $p(\mathcal{T}_i^p|[\mathcal{G}_X = 1], \tau)$ and $p(\mathcal{T}_i^p|[\mathcal{G}_X = 0], \tau)$ must sum up to 1.

$$p(\mathcal{T}_i^p|[\mathcal{G}_X = 1], \tau) \tag{7}$$
$$= \begin{cases} 1/d_{max} & \text{if } \mathcal{T}_i^p < d_X - \epsilon \\ (1 - d_X/d_{max})/\epsilon & \text{if } d_X - \epsilon \leq \mathcal{T}_i^p \leq d_X \\ 0 & \text{if } \mathcal{T}_i^p > d_X \end{cases}$$

$$p(\mathcal{T}_i^p|[\mathcal{G}_X = 0], \tau) = 1/d_{max} \tag{8}$$

To get an intuitive idea of the RIM camera model, imagine we have a single pixel RIM camera, with the depth detection standard deviation $\sigma = 0.3m$ and maximum detection range of $8m$. If the current sensor readout is $5.0m$, according to our RIM sensor model, we can plot out the space occupancy probability $p(\mathcal{G}_X|\mathcal{O}, \mathcal{D}, \tau)$ along the viewing ray as in Fig. 7, given Eq. 3-4 & 6-8. This means the object is most likely existing at $5m$, the observed depth region. Regions in front of it should be free of any object and visible up to the camera. Regions behind $5m$ remains total uncertainty, 0.5, because we have no idea whether there is matter behind the surface or not. The peak falls smoothly on both directions, because of the limited sensor precision. This plot is consistent with the depth sensor models described in other literatures such as [5, 22].

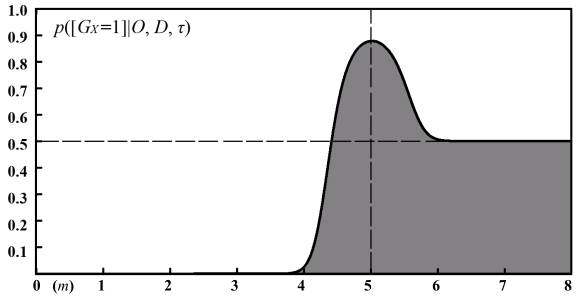

Figure 7. Space occupancy probability $p(\mathcal{G}_X|\mathcal{O}, \mathcal{D}, \tau)$ at certain distances given the RIM camera readout of $5.0m$. It is a longitudinal cut of what probabilities look like on one viewing ray in the grid.

5. Experiment and Result

We acquire two sets of data to test our proposed calibration and heterogeneous sensor framework. Without losing generality, for the camcorders and RIM cameras, we use Canon HG10 and Swiss Ranger 3100 respectively. Canon HG10 DV camcorders are set to run at 25 fps with an image resolution of 1920×1080 pixels. Swiss Ranger 3100 are set to run at 5 fps with an image resolution of 176×144 pixels. The dataset specifications are listed below. For SENSOR NETWORK 2, in order to prevent the interference between multiple RIM cameras, their modulation frequency are manually set at 19MHz, 20MHz and 21MHz respectively. Although this setting will affect the maximum detection depth of each camera, the minimal range $7.1m$ [14] is still beyond our reconstruction volume range, $6m$. Both datasets use a occupancy volume.

	Canon HG10	SR 3100	volume size
Sensor Network I	3	1	$128 \times 256 \times 128$
Sensor Network II	6	3	$128 \times 128 \times 128$

5.1. Depth Calibration Evaluation

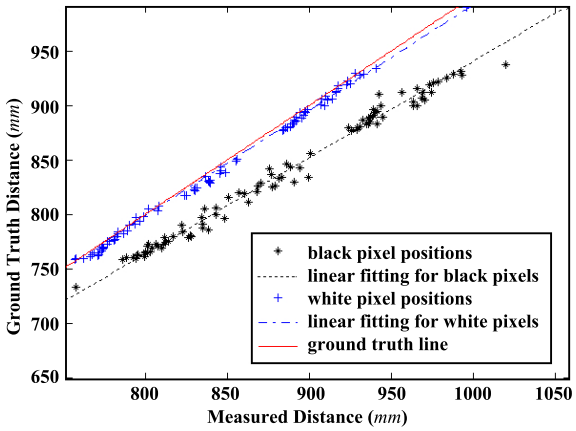

Figure 8. Distance Linear Fitting result. Since the data points are close to one other, this figure is a zoomed in view, and not all the fitting data is shown here. Best viewed in color.

We show the depth calibration procedure with **Sensor Network I**, 3 camcorder and 1 RIM camera setup. 24 checkerboard poses seen by all 4 cameras are used for the geometric calibration. After using Bouguet's MATLAB calibration toolbox [3] to recover the intrinsic and extrinsic parameters for each camera, we perform a bundle adjustment to globally optimize the camera parameters and checkerboard poses. The resulting average pixel re-projection er-

ror is within 0.69 pixels. Since we now know the checkerboard absolute poses, we can compute certain points on the checkerboard to the SR 3100 as the ground truth, and compare the distance values with the SR 3100 sensor measurement. Specifically, for every black square and white square of the checkerboard pattern, we find its center point position's measured distance against the ground truth. As shown in Fig. 8, the measurements do have a linear deviation from the ground truth. If we analyze the black or white patterns of the checkerboard separately, the distance measure is also affected by the received infrared intensity: the darker the pixel intensity is, the less accurate the measurement is. So we fit lines to white and black pixel measurement separately. Thus we get two mapping functions from the measured distance to the ground truth as below. The computed standard deviations show the uncertainty of the measurement, which are also used as the standard deviation in $\mathcal{N}(\mathcal{T}_i^p, \sigma)$ of the **depth measurement term** of Eq. 6.

$$d_{correct} = a \cdot d_{measure} + b, \quad \sigma$$
$$a_{black} = 0.8823, \quad b_{black} = 55.27, \quad \sigma_{black} = 9.131$$
$$a_{white} = 0.9666, \quad b_{white} = 22.70, \quad \sigma_{white} = 6.168$$

These are our depth calibration functions. They are used to correct the depth measurements in the 3D reconstruction later on. Given a certain pixel intensity, we obtain the specific depth correction line parameters a_{pixel} and b_{pixel} by linear interpolation between (a_{black}, a_{white}) and (b_{black}, b_{white}) respectively. It is a very simple solution, but its effectiveness is shown with reconstruction results in the next section. However, with our geometrically calibrated camcorders, more delicate analysis can be performed to explicitly model the distance measurement relationship to intensity changes etc., similar to [19].

5.2. Sensor Network I result

We have two static reconstructions with this 4 camera setup: an office chair with two boxes and a sitting person. The output of the alrogithm is a probabilistic volume, for visualization purpose, the volume surfaces are extracted at an arbitrary iso-probability of 87%, and the results are shown in Fig. 9 and Fig. 10. The reconstructions from our proposed framework preserve detailed concavity and significantly improve the quality of the result from the 3 camcorder only (the probabilistic visual hull). More delicate surface extraction schemes can be applied to get better object shapes, but this is beyond the scope of this paper.

In order to evaluate the depth calibration, we compute another two volumes with only the SR3100 camera turned on, one with the depth correction, and the other without it. Then together with the volume of 3-Camcorder (the probabilistic visual hull), we extract three horizontal slices

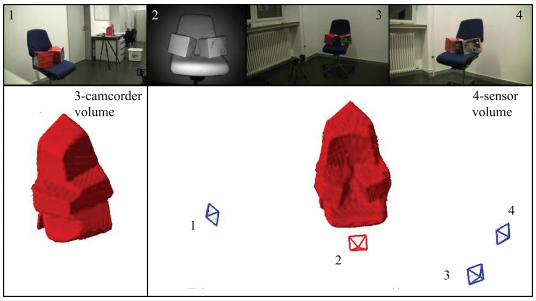

Figure 9. An office chair with two boxes. Top: the four camera views Bottom: 3-camcorder probabilistic visual hull and 4-camera fusion result with our proposed algorithm. The calibrated camera configuration is also shown here, with #2 the SR3100, and 1, 3 and 4 the Canon HG10.

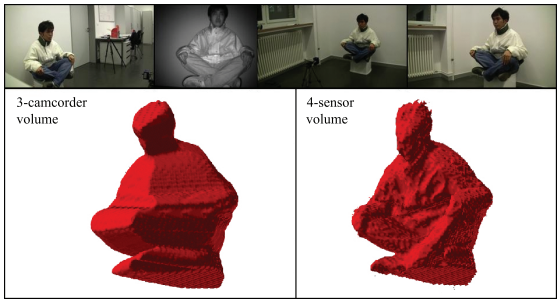

Figure 10. A sitting person. The same configuration with Fig. 9.

at the same height level (the level of the head), as shown in Fig. 11, the intensity denotes the occupancy probability with 0 being black and 1 being white. In theory, the actual object surface should be tangent to the visual hull of the object [16]. After we overlay the visual hull slice to the two SR3100 slices, we see that of the two SR3100 volumes, the one without depth calibration is not tangent with the 3-camcorder visual hull surface, which figuratively will carve away voxels that are actually on the reconstruction object. However, the volume after depth calibration is tangent to the visual hull, demonstrate the necessity and effectiveness of the our depth calibration procedure. The roughly horizontal white lines in Fig. 11 (b) and (c) are the wall position at the back of the person. The thickness reflects the sensor measurement uncertainty, similar to the peak in Fig. 7.

5.3. Sensor Network II result

For this 9 camera network, we also have two reconstructions: a person with a rubber ball, and a crowd of 5 people. The number of cameras in use is not designed on purpose, instead is based on the number of sensors available. Ad-

mittedly though, more detailed information can be obtained with more sensors, and it really helps in challanging cases such as very cluttered scenes. The results are shown in Fig. 9 and Fig. 10 respectively. The camera calibration procedures are the same as the previous dataset. And the recovered camera poses are shown in Fig. 12, with red cones denoting three SR3100. The reconstructed ball in Fig. 12 has a diameter of 60cm, which is pretty close to the actual value is 57.06cm given the low volume resolution. This again shows the power of our depth calibration. A more challenging example is Fig. 13, where 5 people are highly clustered in the space. Without the depth information to recover the concavities the visual hull would fail the reconstruction task. One thing to note is that the missing forearms are sub-voxel size. They can be recovered if we increase volume resolution at those places.

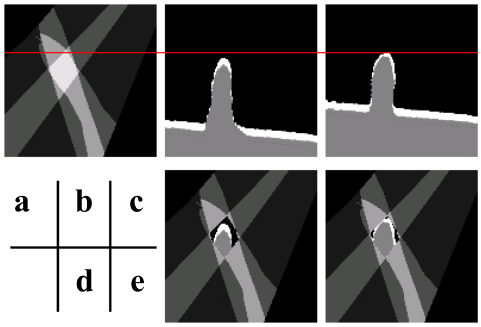

Figure 11. Horizontal slices at the head level of the sitting person data. Cameras are looking downwards from the top. (a) 3-camcorder volume. (b) SR3100 volume without depth correction. (c) SR3100 volume without depth correction. (d) overlays (a) to (b), there is a big gap between the two. (e) overlays (a) to (c), they are tangent. The red line on the first row shows the depth measure difference before and after depth calibration. From (d) and (e), it is shown that our depth correction gives more accurate front most surface, which should be tangent to the visual cone.

6. Discussion

In this paper, we propose a new heterogeneous sensor network of camcorders and RIM cameras in multi-view 3D object reconstruction. To achieve more accurate distance measurements, we carry out a new RIM camera depth calibration method as a simple extension of the conventional camera geometric calibration process. We then propose a novel probabilistic sensor fusion framework to robustly relate camcorder silhouette cues and RIM camera depth images together, and improve the reconstruction quality significantly comparing with the result using either type of sensor alone. RIM cameras are thus shown for the first time to be a very promising new type of sensor for accurate multi-view 3D reconstruction, besides its proposed usage in object detection, tracking etc. For camcorders, similar to [11], no explicite silhouette extraction is needed. More importantly, our sensor fusion framework is general enough and not limited to a silhouette cues or depth images, but also to disparity maps of stereo camera pairs or 3D point clouds of LIDAR sensors etc., as long as the proper sensor model is provided. Also, using our camcorder-RIM camera platform, similar to our depth calibration process, with the guidance from the geometrically calibrated camcorders, more delicate experiments can be carried out to analyze RIM camera's impulse-response properties, such as depth measure variation with respect to infrared light incident angle or material reflectance of the object described in [15]. Finally, consider computation time to our volume framework, most of the computation can be parallelized on GPU. Also given

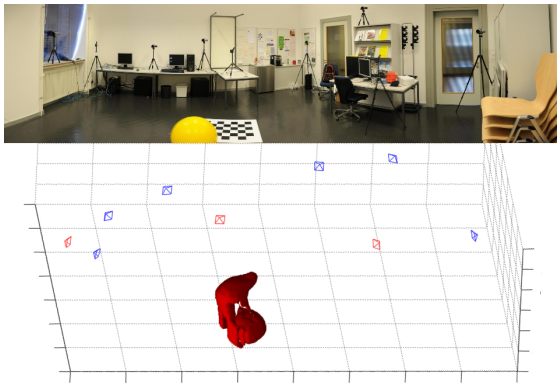

Figure 12. Top: the camera settings. Bottom: the reconstruction of a person with a rubber ball. Best viewed in color.

Figure 13. The reconstruction of the densely populated scene from all 9 sensors with concavity and details. Visual hull fails in this case, resulting an indistinguishable blob.

the high frame rate of both the camcorders and RIM cameras, dynamic scenes can be recovered in real-time.

Acknowledgments: We would like to thank Rolf Adelsberger, Prof. Markus Gross, Tobias Kohoutek and Prof. Hilmar Ingensand for resource support and technical discussion. This work was partially supported by David and Lucille Packard Foundation Fellowship, and NSF Career award IIS-0237533.

References

[1] C. L. Bajaj, F. Bernardini, and G. Xu. Automatic reconstruction of surfaces and scalar fields from 3D scans. *Computer Graphics*, 1995.

[2] B. Baumgart. Geometric modeling for computer vision. *PhD thesis, CS Dept, Stanford U.*, 1974.

[3] J.-Y. Bouguet. Camera calibration toolbox for matlab. *http://www.vision.caltech.edu/bouguetj/calib*$_d$*oc/*.

[4] G. Casciola, D. Lazzaro, L. B. Montefusco, and S. Morigi. Fast surface reconstruction and hole filling using positive definite radial basis functions. *Numerical Algorithms*, 2005.

[5] C. Coué. Modèle bayésien pour l'analyse multimodale d'environnements dynamiques et encombrés : Application à l'assistance à la conduite en milieu urbain. *Dissertation to doctor of Sciences Institut National Polytechnique De Grenoble*, 2003.

[6] B. Curless and M. Levoy. A volumetric method for building complex models from range images. *Computer Graphics*, 1996.

[7] J. Davis, S.R. Marshner, M. Garr, and M. Levoy. Filling holes in complex surfaces using volumetric diffusion. *3DPVT*, 2001.

[8] A. Elfes. Occupancy grids: a probabilistic framework for robot perception and navigation. *Dissertation to doctor of Sciences CMU*, 1989.

[9] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. *ECCV*, 2000.

[10] J.-S. Franco and E. Boyer. Exact polyhedral visual hulls. *BMVC*, 2003.

[11] J.-S. Franco and E. Boyer. Fusion of multi-view silhouette cues using a space occupancy grid. *ICCV*, 2005.

[12] W. Grimson and C. Stauffer. Adaptive background mixture models for real-time tracking. *CVPR*, 1999.

[13] A. Hilton, A. Stoddart, J. Illingworth, and T. Windeatt. Implicit surface based geometric fusion. *CVIU*, 1998.

[14] MESA Imaging. Swiss ranger 3000 help document, 1.0.8.x, miniature 3d time of flight camera. *support@swissranger.ch*.

[15] T. Kahlmann. Range imaging metrology: Investigation, calibration and development. *Dissertation to doctor of Sciences ETH Zurich*, 2007.

[16] A. Laurentini. The visual hull concept for silhouette-based image understanding. *PAMI*, 1994.

[17] S. Lazebnik, Y. Furukawa, and J. Ponce. Projective visual hulls. *IJCV*, 2007.

[18] M. Li, H. Schirmacher, M. Magnor, and H.-P. Seidel. Combining stereo and visual hull information for online reconstruction and rendering of dynamic scenes. *Computer Graphics*, 1996.

[19] M. Lindner and A. Kolb. Calibration of the intensity-related distance error of the pmd tof-camera. *SPIE*, 2007.

[20] R. Sablatnig M. Kampel and S. Tosovic. Fusion of surface and volume data. *OAGM*, 2002.

[21] D. Margaritis and S. Thrun. Learning to locate an object in 3d space froma sequence of camera images. *ICML*, 1998.

[22] K. Pathak, A. Birk, J. Poppinga, and S. Schwertfeger. 3d forward sensor modeling and application to occupancy grid based sensor fusion. *IROS*, 2007.

[23] R. Sablatnig, S. Tosovic, and M. Kampel. Combining shape from silhouette and shape from structured light for volume estimation of archaeological vessels. *ICPR*, 2002.

[24] D. Snow, P. Viola, and R. Zabih. Exact voxel occupancy with graph cuts. *CVPR*, 2000.

[25] R. Szeliski. apid octree construction from image sequences. *Computer Vision, Graphics and Image Processing*, 1993.

[26] R. Whitaker. A level-set approach to 3d reconstruction from range data. *IJCV*, 2004.

[27] Y. Yemez and C.J. Wetherilta. A volumetric fusion technique for surface reconstruction from silhouettes and range data. *CVIU*, 2007.

[28] Z. Zhang. A flexible new technique for camera calibration. *PAMI*, 2000.