A Unified Approach to Calibrate a Network of Camcorders and ToF cameras

Li Guan Marc Pollefeys

lguan@cs.unc.edu marc.pollefeys@inf.ethz.ch

UNC-Chapel Hill, USA.
ETH-Zürich, Switzerland.

Abstract. In this paper, we propose a unified calibration technique for a heterogeneous sensor network of video camcorders and Time-of-Flight (ToF) cameras. By moving a spherical calibration target around the commonly observed scene, we can robustly and conveniently extract the sphere centers in the observed images and recover the geometric extrinsics for both types of sensors. The approach is then evaluated with a real dataset of two HD camcorders and two ToF cameras, and 3D shapes are reconstructed from this calibrated system. The main contributions are: (1) We reveal the fact that the frontmost sphere surface point to the ToF camera center is always highlighted, and use this idea to extract sphere centers in the ToF camera images; (2) We propose a unified calibration scheme in spite of the heterogeneity of the sensors. After the calibration, this multi-modal sensor network thus becomes powerful to generate high-quality 3D shapes efficiently.

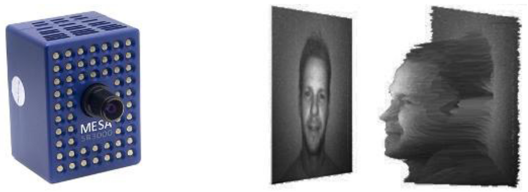

Fig. 1. Left: A ToF camera, Swiss Range 3000. Right: Typical output from the sensor: an intensity (amplitude) image and a 2.5D depth image. (Courtesy of MESA imaging.)

1 Introduction

With the new advances in vision sensor technologies, Range Imaging (RIM) cameras based on Time-of-Flight (ToF) principles are becoming more and more popular within the past few years. Typically, a ToF camera emits a modulated optical radiation field in the infra-red spectrum. This signal is diffusely backscattered by the scene and detected by the camera. Every CMOS/CCD pixel is able to demodulate the signal and detect its phase, which is proportional to the distance of the reflecting object. Although, currently most of these ToF cameras do not have high image resolution (e.g. 176×144 for Swiss Ranger 3000, as shown in Fig. 1), they can generate a 2.5D *depth image* together with an *intensity image* (an *amplitude image* in some literatures) at a frame rate up to $50fps$, which is far beyond the throughput of traditional depth sensors, such as LIDAR. They thus have an enormous potential in a wide range of applications including 3D object reconstruction, especially real-time dynamic scene reconstruction.

A powerful multi-modal sensor network for 3D reconstruction is proposed in [1], where a few ToF cameras and video camcorders work together. After a traditional checkerboard calibration [2,3] of the network, the complementary 2.5D depth information from the ToF cameras and high-res silhouette information from the camcorders are fused together to recover a robust probabilistic 3D representation of an object. The colored textures from the camcorders can be further applied in the final rendering stage. Despite the robustness of the sensor fusion, the reconstruction quality still could be improved. One of the main issues is the geometric calibration of the network.

1.1 Related work on vision sensor network calibration

All the literatures related to ToF camera calibration are focused on the single camera extrinsics calibration or depth calibration [4,5]. For camera/camcoder, contrarily, multi-view calibration has been well studied over the past decades.

The most common multi-view calibration technique for pure camera/camcorder network is to use a black & white checkerboard pattern as a calibration target, and use Zhang's method [2,3] followed by a global bundle adjustment to recover the intrinsics as well as the extrinsics of all the cameras. To extend this approach for our camcorder & ToF camera network, the first problem is the low ToF camera image resolution. The extracted checkerboard corners are reported to be not accurate [1]. The second drawback is that the planar checkerboard pattern cannot always be guaranteed to be seen by all the cameras at the same time, such as the dataset used in our paper in Section 3. Therefore, in order to recover the complete camera configuration, one has to first calibrate neighboring cameras that observe the common checkerboard poses and then do piece-wise bundle adjustments to link local configurations together. It is really tedious labor work. Moreover, if in some extreme cases, two neighboring cameras do not witness sufficient common checkerboard poses, the global linkage would never be accomplished.

An alternative solution for multi-view camera/camcorder calibration is to use a single 3D laser point [6] to recover the parameters up to the global scale ambiguity, or a rigid grid of 3D points [7] to recover the full Euclidean space. In these cases, all the cameras can see the calibration target simultaneously, thus solve the second problem of the previous planar calibration target approach.

But because of the low image resolution, the detection of this type of calibration targets is not always robust in the ToF camera images.

In this paper, we follow the second type of approach, but instead of using a moving laser dot or a grid of rigidly connected points, we use a moving sphere of an unknown radius as our calibration target. Literatures [8,9,10] also propose to use a spherical calibration target for a camera/camcorder network. The main difference is that they use the complete sphere contour information constrained by the absolute conic. However, due to the low resolution of the ToF camera images, the sphere contour extraction is risky for our setup.

Whereas we exploit the fact that the sphere center can be robustly extracted not only from the high-res camcorder frames, but also the low-res ToF frames, based on our original observation that *in a ToF camera intensity image, a sphere center is always highlighted*. The highlight is due to the ToF camera active sensing mechanism, the surface reflectivity of the sphere and the spherical surface normal direction property. Real images from an SR3100 ToF camera are shown in Fig. 3 and Fig. 5. After the sphere center locations are extracted, we can perform a bundle adjustment similar to [6] recovering the global extrinsic camera poses and sphere center 3D locations. Given the depth information from the ToF cameras, the global scale and the full Euclidean space can be recovered.

The paper is organized as follows: In Section 2, we introduce our method to calibrate the multi-modal network, including the sphere center extraction, bundle adjustment and sphere radius & global scale recovery. In Section 3, we validate the proposed method with a real example of a two-camcorder-two-ToF-camera network. A probabilistic volumetric 3D object reconstruction is also given to demonstrate the power of such a heterogenous sensor network. Finally in Section 4, we discuss some remaining issues, further directions to go, and summarize the paper.

2 Unified calibration scheme

We move around a sphere with an unknown radius in the common viewing space of the sensor network. Let's assume the intrinsic parameters of the sensors are known, either from factory specification of the cameras or through a pre-calibration [3]. Thus the image radial distortion can be removed. By extracting the sphere center from the synchronized video frames from all sensors, we can perform a bundle adjustment over sensor extrinsic parameters and sphere 3D locations, and solve the system up to a scale factor. Then, we can recover the global scale ratio using the ToF sensor depth measurement and our already computed 3D sphere centers locations in the similarity space. Notice that this scale ratio is not taken into account during the geometry bundle adjustment, because the ToF camera depth measurement sometimes can be very noisy. We therefore do not rely on it to be an effective optimization constraint, but recover it in a separate step.

2.1 Sphere center detection

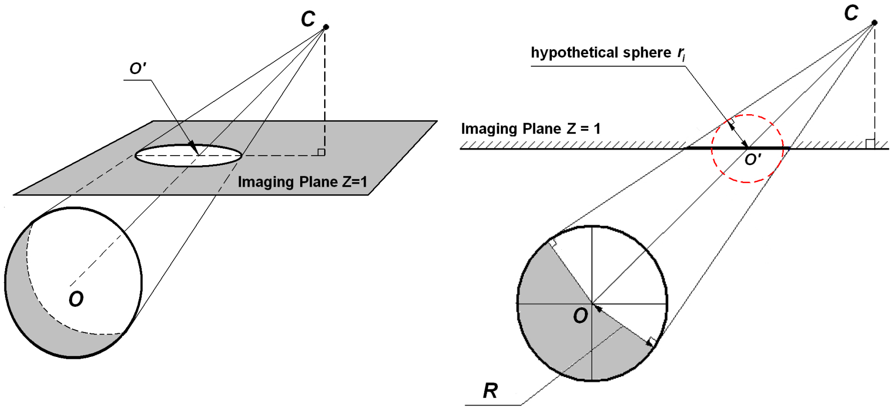

Fig. 2. Left: A sphere projection on to the image of a camera centered at C is in general an ellipse, due to the projective distortion. The sphere center's projection O' is also not at the ellipse center. Right: a 2D side view of the same configuration. The "hypothetical sphere" located at $(x_i, y_i, 1)^\mathsf{T}$ is shown as the red dotted circle with radius r_i.

Camera/camcorder After the radial distortion is removed, the 2D image of a sphere is an ellipse [8,9,10], as illustrated in Fig. 2. Hough transform is applied for robust ellipse detection. In general, an ellipse is defined by five parameters. So the Hough space is 5D. However, since we know the camera principal point, we can unambiguously define an ellipse in an image i by the sphere radius R and the viewing ray from the camera optical center to the sphere center, namely vector $\langle x_i, y_i, 1\rangle$. But since R is unknown, we need an alternative way to describe the radius. Given that the intrinsics are known, for every image i, we can introduce a "*hypothetical sphere*" located at $(x_i, y_i, 1)^\mathsf{T}$ – one can think of a sphere located on the $Z = 1$ plane in the 3D space – with a varying radius r_i, such that this new sphere results in the same elliptical image projection as our *real* radius R sphere, as shown in the right plot of Fig. 2. Therefore, we actually have only an (x_i, y_i, r_i) 3D Hough space, which makes the computation easier. Practically, we use the edgeness cue and the color of the sphere to guide our Hough transform. To further exploit the temporal consistency between neighboring image frames, given $(x_{i-1}, y_{i-1}, r_{i-1})$ in frame $i - 1$, we apply a simple tracking scheme to constrain the local search window for detection in frame i.

ToF camera A sphere center in a ToF camera image should not be extracted in the same way as a camera/camcorder, because of the extremely low image resolution and relatively bad sphere edge contrast. However, thanks to the ToF camera active sensing mechanism, we have an even simpler way to find a sphere center.

Most of the current ToF cameras (e.g. Swiss Rangers, PMD sensors and Canesta cameras) have LEDs evenly distributed around the camera lens. The

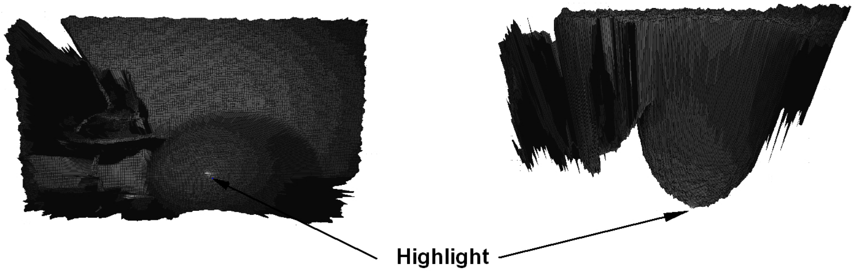

Fig. 3. Left: A typical intensity image of a Swiss Ranger 3100 mapped on to its depth mesh. The highlight is due to most of active light reflection in a local region. Right: A top view of the same depth mesh. The highlight is along the viewing ray through the camera optical center and sphere center.

active light can thus be think of as from a single virtual "*point light source*" located at the camera optical center. Therefore, assume a Lambertian sphere surface, which has the property that *the witnessed brightness in an image depends only on the angle between the surface normal and the light source*, the brightest pixel in the intensity image is the one that lies on the line connecting the camera optical center (the virtual "*point light source*" position) and the sphere center, which by definition is the sphere center's projection in the image. In fact, the Lambertian assumption is not necessary, since the specularity of the surface even strengthens the highlight, due to the overlapping virtual light source with the camera optical center. Both Fig. 3 and Fig. 5 show this phenomenon in a real SR 3100 image.

Some might point out that strictly speaking the assumption that we have a virtual "*point light source*" does not hold, because the light source is not a point (but a network of LEDs), the non-perfect shape and texture of an actual spherical object, and aliasing effects. But we would like to thank instead of blame the low resolution in our case that the highlight blob we observe is so small that only with in the size of a pixel and our assumption still holds, as long as the sphere is far away enough from the imaging plane and the relative sphere radius is much bigger than the dimension of the light source array.

Also some might have noticed that the depth image from the ToF cameras as well gives us some hint to the sphere center's image location, but there are at least two reasons that we do not compute from it. First of all, the closest point on the sphere depth mesh is usually NOT the sphere center location, unless the sphere center goes through the principal axis. Thus the computation would not be as easy and direct as our following proposal. Secondly, the depth image is noisy, as shown in Fig. 3.

Here comes our actual proposal: To recover the sphere center's image location in each ToF frame, we detect and track the highlight in the intensity image only. A paraboloid is fitted to achieve sub-pixel accuracy, given the intensity values around the found maximum location.

2.2 Recover the extrinsics via bundle adjustment

Given the intrinsics and sphere center's image locations in the synchronized frames from all the views, we can now perform a global bundle adjustment similar to [6] to recover the camera poses and sphere 3D locations. The intrinsics of the video camcoders are pre-computed using [3]'s method. The intrinsics of the ToF cameras are obtained from the factory manual. But we can also put the intrinsics into the bundle for a further refinement. Due to the heterogeneity of the sensors, namely the image resolutions are very different and the sphere center extraction methods we have just described are very different (The uncertainty of the camcorder Hough transform is related with the sphere boundary extraction, intrinsic optical center computation, radial distortion correction and Hough space resolution; The uncertainty of the ToF camera sphere center extraction is related with image noise and motion blur.), to minimize the algebraic error (pixel re-projection error) is meaningless. Therefore, we define the bundle adjustment error metric to be the angular re-projection error [11,12,13], i.e. the angle θ between the observed ray $\boldsymbol{x}$ and the re-projection ray $\boldsymbol{r}$:

$$f(\boldsymbol{X}) = |\tan(\theta)| = \left| \frac{\boldsymbol{x} \times \boldsymbol{r}}{\boldsymbol{x}^{\mathsf{T}} \boldsymbol{r}} \right| . \tag{1}$$

This overcomes the image resolution difference. Since it is hard to model the methodological distinction between the two sphere center extraction approaches, for now we assume the two methods have equal uncertainties, thus ignore this issue in the rest of the paper.

2.3 Recover the sphere radius R and global scale S

After the bundle adjustment, we can compute the relative distance from each sphere center 3D location to the camcorder optical centers. And since we have r_i from the Hough transform, for each camcorder image i, we can compute the 3D sphere radius R_i by similar triangle analysis, as shown in the right plot of Fig. 2. For the real sphere radius R, we just take the mean of all R_i s. Notice that R is still in the similarity space, but not the metric radius of our calibration target.

To recover the global scale S, we read out the depth measurements D_i at the sphere center's pixel position from the ToF depth images. And suppose the relative distance from the sphere center to the ToF camera optical center is d_i, we have the expression below. Detailed implementation is described in Section 3.4 with a real dataset.

$$D_i = (d_i - R) \cdot S. \tag{2}$$

3 Result and evaluation

3.1 Experiment setup

In this section, we describe a real dataset consists of two Canon HG10 camcorders and two Swiss Ranger 3100 ToF cameras. The camcorders are set to run at 25 fps with an image resolution of 1920×1080 pixels. The Swiss Rangers are set to run at 20 fps with an image resolution of 176×144 pixels. Although many delicate approaches could be applied, we simply synchronize our four views by temporally sub-sampling the frames at 5 fps. Our calibration target is a yellow gymnastic ball. The four sensors locate on a rough circle, looking inward at a common free space. The two ToF cameras are set at different modulation frequencies, i.e. 19 MHz and 20 MHz, so that the active lights are not interfering with each other. And this gives a minimum unambiguous depth range of 7.1 meters, which well satisfies our current indoor environment.

3.2 Sphere center extraction

The sphere centers are extracted using the described methods in the previous section. The camcorder image Hough transform is illustrated in Fig. 4. And the ToF camera sphere center highlight is shown in Fig. 5. The extracted sphere center locations for all four views are shown as green dots in Fig. 6. In order to get an unbiased and robust calibration, we intensionally move the sphere to sample the 3D space as uniformly as possible.

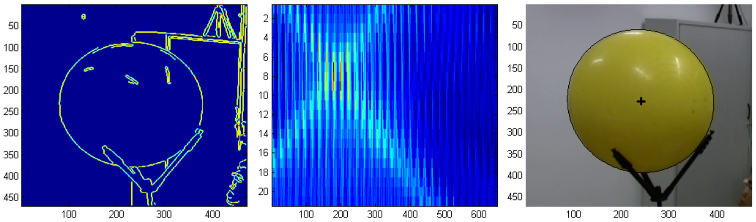

Fig. 4. Camcorder ellipse extraction. Only cropped images are shown. **Best viewed in color.** Left: A thresholded gradient magnitude image. Middle: 2D projection of the (x_i, y_i, r_i) 3D Hough space. A single solution is found, at the crossing. Right: The recovered ellipse and the sphere center overlaid on the original image.

3.3 Bundle adjustment evaluation

After the bundle adjustment, the recovered camera configuration and 3D sphere center locations are shown in the left plot of Fig. 7. The plot on the right is the re-projection error statistics in the box-and-whisker diagram from our bundle adjustment result. A different way to evaluate our recovered poses is to project the camera centers to different camera views, as shown in the yellow crosses in Fig. 6. One can see that the projected camera centers well overlay their images from a different view as expected.

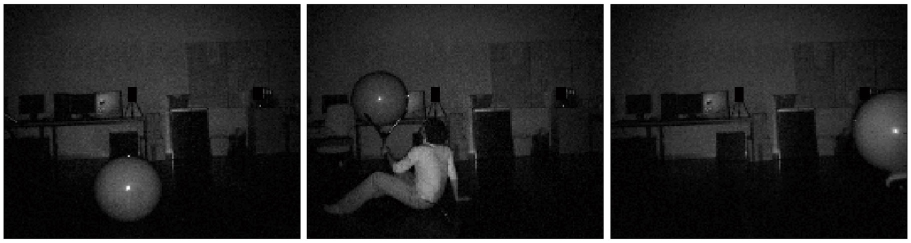

Fig. 5. ToF camera Swiss Ranger 3100 intensity image sphere center highlights. To detect the highlight robustly and automatically, we apply a simple Region of Interest (ROI) tracking method.

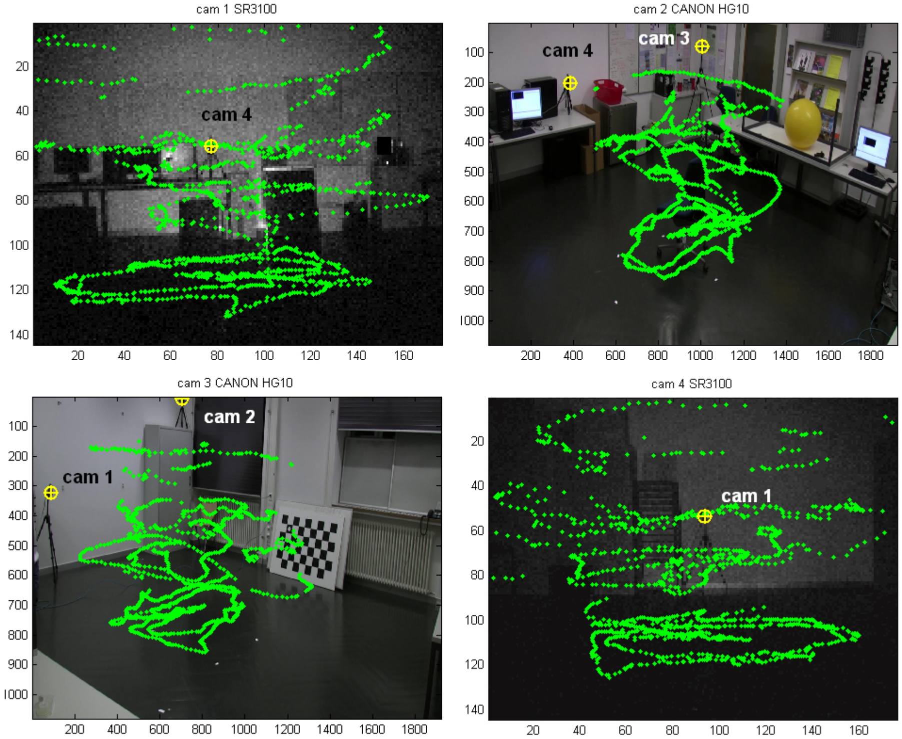

Fig. 6. The extracted sphere center's image locations are shown in green dots. The sphere samples cover the 3D space as much as possible. After the bundle adjustment, the recovered camera centers are re-projected to the images as yellow crosses. They overlap very well with the camera image, showing that the system is well calibrated.

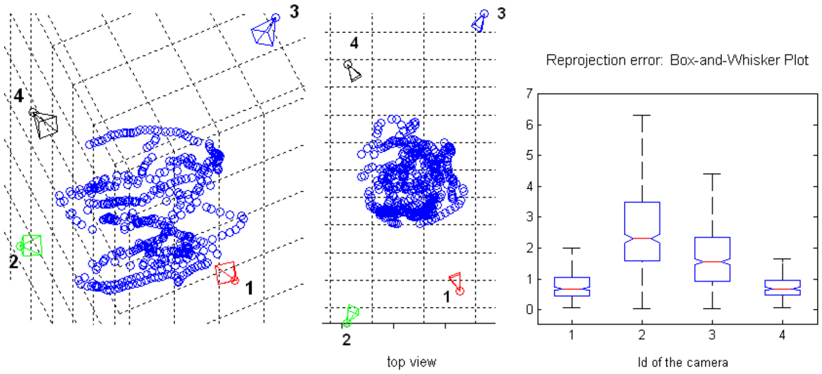

Fig. 7. Left: The recovered sphere center's image locations for all four views. Right: Image re-projection error statistics. Cam #1 and #4 are ToF cameras. Cam #2 and #3 are video camcorders. The re-projection errors for the ToF cameras are significantly smaller than those of the video camcorders, whose image resolution is much higher. This shows our bundle adjustment does not have a bias to the resolution difference.

For completeness, the physical projection matrices of the four cameras in the setup are listed in Tab. 1 (sensor internal and external parameters) and the full dataset used for reconstruction in Section 3.5 is available upon email request.

Table 1. Recovered camera projection matrices with our method.

Cam #1	-69.098	123.3147	152.3223	340.4835
	-176.7834	-83.9725	46.8359	103.5043
	-0.0888	-0.3609	0.9284	1.2298
Cam #2	-799.49	1401.3	286.95	2048.2
	-1418.8	-360.37	-123.77	416.72
	-0.56033	0.18502	0.80743	1.7704
Cam #3	380.59	-1900.0	115.72	-1140.6
	-1716.4	-704.98	-151.45	296.97
	-0.14273	-0.54063	-0.82907	-1.0773
Cam #4	25.9357	-129.1248	-167.5592	-304.4062
	-209.1743	-37.0983	-45.5719	31.5774
	-0.3153	0.42896	-0.84651	-0.42014

We also actually have tried to calibrate the system with a planar checkerboard pattern for comparison, but fail to link the camera pairs (#1, #2) and (#3,#4) together. Because as we see in Fig. 7, view #1 is almost opposite to view #3, so as #2 to #4, which unfortunately is one of the extreme cases having been discussed in Section 1.1. This again shows the advantage of our calibration approach.

3.4 Sphere radius and global scale recovery

Given the recovered 3D structure and camera poses, and the hypothetical sphere radius r_is, we can compute the mean radius $R = 0.0248$, simply by similar triangle inference, as shown in the right plot of Fig. 2. Note again that R is not the true sphere radius, but in our recovered similarity space.

To recover the absolute scale S, we can re-write Eq. 2 as a minimization problem:

$$arg \min_S \sum_i |D_i + R \cdot S - d_i \cdot S| \, . \tag{3}$$

Given our dataset, we solve $S = 11.3886$, and the absolute sphere radius $R' = R \cdot S = 0.2824\ m$. We measure the sphere circumference to be $1.7925\ m$, namely the measured sphere radius is $0.2853\ m$, which is very close to our computation.

3.5 3D reconstruction application

Once the sensor network configuration is fully recovered, we can perform the multi-view reconstruction. We model the 3D space as a probabilistic occupancy grid. The 3D space is represented by a 256^3 probability volume. Each voxel is assigned the posterior probability to be occupied by the person, given the sensors' observations. Using the Bayesian sensor fusion graphical model introduced in [1] and the MATLAB source code at [14], a robust 3D probabilistic and volumetric shape estimation can be computed. Surface model is then obtained by thresholding the volume at 0.875 (manually tuned for visual quality), as shown in Fig. 8. Despite the limited number of views, we are able to recover details on the surface, thanks to the complementary depth information from the ToF sensors and silhouette information from the camcorders. As a comparison, if we only have camera silhouette information, although we still can perform the same volume computation with the method from [14], and the output can be thought of as a robust four-view "probabilistic visual hull". Self-occlusions introduce several "ghost" regions, especially at the person's legs. This actually shows the true power of such multi-modal sensor networks over just single-modal camcorder clusters and the silhouette information alone. Also a two-camcorder-two-ToF-camera network in our dataset arrangement (two ToF sensors facing each other; two camcorders facing each other, orthogonally to the ToF sensors) provides a minimum number of sensors required for detailed surface shape and sufficient texture map. Another illustration of the volumes are shown in Fig. 9.

4 Discussion and summary

From our calibrated four-sensor setup, we show the possibility of using two opposite-posing ToF cameras for detailed geometry reconstruction and other two opposite-posing video cameras for extra guidance and more importantly complete texture maps. Since all the sensors in this setup can run in real-time, this setup can be thought of as a simple dynamic scene reconstruction configuration,

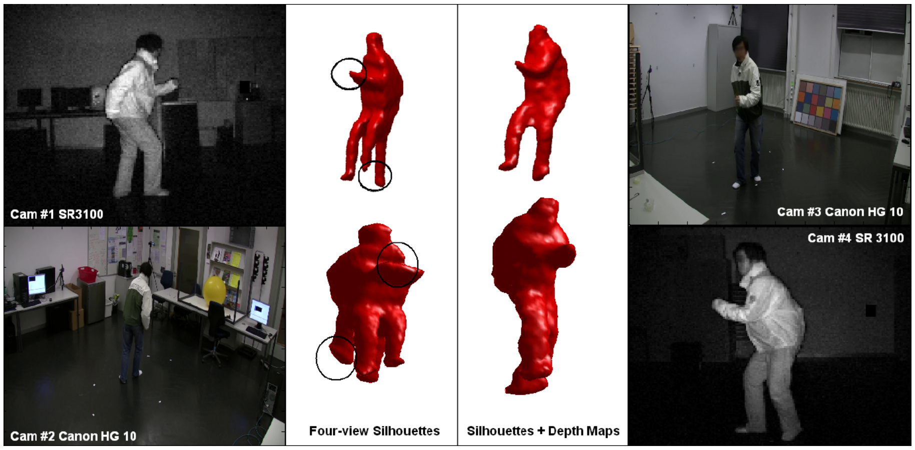

Fig. 8. 3D reconstruction from the calibrated heterogeneous sensor network. All the views are listed on the sides. Detailed reconstruction is achieved by combining the multi-modal visual cues, namely the depth map from the ToF cameras and silhouette information from the camcorders. The shape is qualitatively better than the "probabilistic visual hull" (see [1] for more detail) from the single-modal silhouette information, specially in regions indicated by the circles. This demonstrates the power of such calibrated multi-modal sensor network in 3D reconstruction applications.

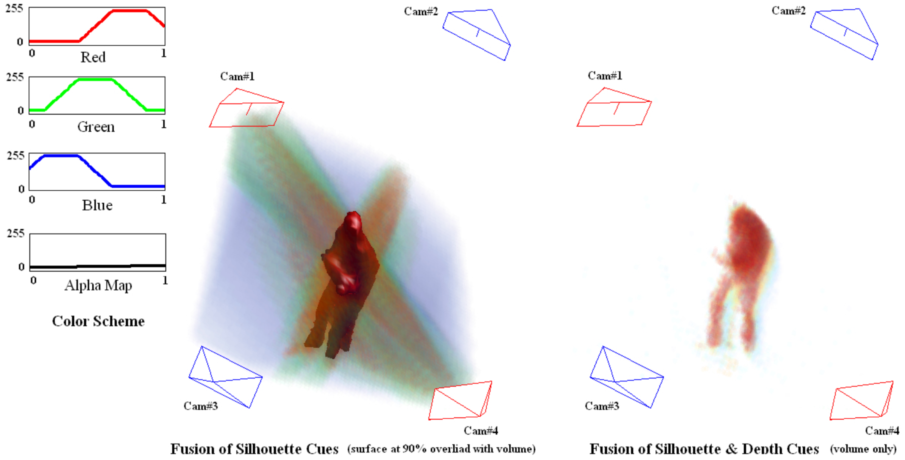
255
0
0
Red
1
255
0
0
Green
1
255
0
0
Blue
1
255
0
0
Alpha Map
1
Color Scheme
Cam#2
Cam#1
Cam#3
Cam#4
Fusion of Silhouette Cues (surface at 90% overlaid with volume)
Cam#2
Cam#1
Cam#3
Cam#4
Fusion of Silhouette & Depth Cues (volume only)

Fig. 9. Visualization of the probabilistic volumes. Left: probabilistic volume of the fusion of only silhouette information from the four views, with the thresholded surface at 0.9 overlaid on top. The probabilistic "viewing cones" and the "ghost legs" are visible. Right: probabilistic volume of the fusion of both silhouette and depth information. Due to the depth information, the "viewing cone" regions have low occupancy probability and thus not visible here. The "ghost leg" ambiguity is also eliminated. **Best viewed in color.**

with very few number of sensors. Another interesting idea worth exploring is the relationship between the ToF camera depth measurement against the active light incident angle to the reflecting surface, a relationship pointed out by [5]. The fact is that after our geometric calibration of the system, the ToF camera images captured during the calibration process should be analyzed further, because of the nice surface normal properties of our spherical calibration target. Future works also include depth calibration refinement, more dataset acquisition and temporal synchronization analysis.

To summarize, in this paper, we propose a new scheme to calibrate heterogeneous camcorder and ToF camera networks with a moving sphere. It overcomes the low resolution ToF camera image issue, and is almost automatic to recover the Euclidean sensor configuration. Both statistical evaluation and real dataset verify the feasibility and power of this calibration approach and this multi-modal sensor network setup.

Acknowledgments: We gratefully thank David and Lucille Packard Foundation Fellowship, and NSF Career award IIS-0237533 for their financial support on this project.

References

1. Guan, L., Franco, J.S., Pollefeys, M.: 3d object reconstruction with heterogeneous sensor data. 3DPVT (2008)
2. Zhang, Z.: A flexible new technique for camera calibration. PAMI (2000)
3. Bouguet, J.Y.: Camera calibration toolbox. (www.vision.caltech.edu/bouguetj)
4. Kahlmann, T.: Range imaging metrology: Investigation, calibration and development. Dissertation to doctor of Sciences ETH Zurich (2007)
5. Lindner, M., Kolb, A., Ringbeck, T.: New insights into the calibration of tof-sensors. CVPR Workshop On Time of Flight Camera based Computer Vision (TOF-CV) (2008)
6. Svoboda, T., Martinec, D., Pajdla, T.: A convenient multi-camera self-calibration for virtual environments. PRESENCE: Teleoperators and Virtual Environments (2005)
7. Uematsu, Y., Teshima, T., Saito, H., Cao, H.: D-calib: Calibration software for multiple cameras system. ICIAP (2007)
8. Agrawal, M., Davis, L.: Complete camera calibration using spheres: A dual-space approach. CVPR (2003)
9. Ying, X., Zha, H.: A novel linear approach to camera calibration from sphere images. ICPR (2006)
10. Zhang, H., Wong, K., Zhang, G.: Camera calibration from images of spheres. PAMI (2007)
11. Oliensis, J.: Exact two-image structure from motion. PAMI (2002)
12. Hartley, R.I., Schaffalitzky, F.: l_∞ minimization in geometric reconstruction problems. CVPR (2004)
13. Ke, Q., Kanade, T.: Quasiconvex optimization for robust geometric reconstruction. ICCV (2005)
14. Guan, L.: Multi-sensor probabilistic volume reconstruction source code in matlab. (`www.cs.unc.edu/~lguan/PVH_MATLABv1.1.zip`)