Fusion of Time of Flight Camera Point Clouds

James Mure-Dubois and Heinz Hügli

Université de Neuchâtel - CH-2000 Neuchâtel
http://www-imt.unine.ch/parlab

Abstract. Recent time of flight cameras deliver range images (2.5D) in real-time, and can be considered as a significant improvement when compared to conventional (2D) cameras. However, the range map produced has only a limited extent, and suffers from occlusions. In this paper, we investigate fusion methods for partially overlapping range images, aiming to address the issues of lateral field of view extension (by combining depth images with parallel view axes) and occlusion removal (by imaging the same scene from different viewpoints).

1 Introduction

Recent *time of flight* (TOF) cameras enable new applications by producing range maps in real-time. Nevertheless, many situations arise where a single range map does not provide enough information. In this paper, we investigate real-time fusion of multiple range images acquired with TOF cameras. This fusion allows to obtain more information on the 3D scene observed. In particular, we will discuss the possibility to extend the field of view and to remove occlusions when using multi-camera networks.

In the following section, the operation principle of a TOF camera is described, and limitations of monocular view are presented. Then, available options to build TOF camera networks are discussed. Section 2 introduces two candidate calibration strategies for such networks : bundle adjustment and surface matching. Those methods are compared in section 3. Although the evaluation is only qualitative, the comparison clearly illustrates the advantages of the surface matching technique. Finally, section 4 presents example applications of multi-camera networks, compatible with real-time operation.

1.1 Operation of a Time of Flight Camera

Time of flight cameras involve active illumination, and deliver range (or depth) data by measuring the time needed for a light signal to travel from the camera light source to the scene and back to the camera sensor, as illustrated in figure 1. Present cameras ([1],[2]) are based on the continuous emission of a periodic signal. The frequency of modulation f of this signal is typically 20MHz. The periodic signal $S(\mathbf{i})$ received at each pixel $(\mathbf{i})$ of the camera sensor is described by its amplitude $A(\mathbf{i})$ and its phase $\varphi(\mathbf{i})$. The range r is directly proportional to the phase. If we note c as the speed of light, we have :

$$S(\mathbf{i}) = A(\mathbf{i}) \cdot e^{\jmath\varphi(\mathbf{i})} \; . \qquad r(\mathbf{i}) = \frac{c}{4\pi\, f} \cdot \varphi(\mathbf{i}) \; . \tag{1}$$

One key advantage of such cameras is their real-time capability : range maps can be delivered at 20 frames per second (fps).

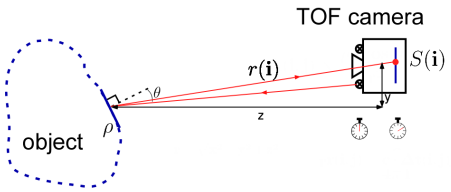
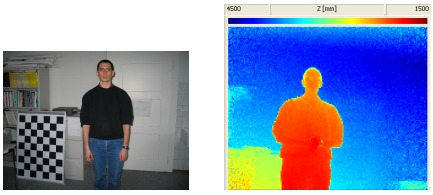


Fig. 1. Time of flight camera - Principle of operation



(a) Color image **(b)** TOF range map $\mathbf{r}\,(i,j)$

Fig. 2. Range map for simple situation (the color image was acquired by a standard camera, and is provided for illustration purposes only).

1.2 Single-Camera Operation, and its Limitations

Our test TOF device, the SwissRanger SR-3000[1] has a 176×144 sensor. Figure 2 illustrates the range map $\mathbf{r}\,(i,j)$ produced by this device for a simple scene. By using a pinhole camera perspective projection model [3], it is possible to transform this range map into a cloud of 3D points. We call f the camera focal length, d_x and d_y the pixel pitch in the x (resp. y) direction, and define normalized pixel coordinates (X_c, Y_c), which are pixel coordinates relative to the position (c_x, c_y) of the optical center on the sensor array. Neglecting lens distortion, the transformation between the range map $\mathbf{r}\,(i,j)$ and 3D coordinates (relative to the camera position) is given by :

$$\begin{aligned} z &= r \cdot \frac{f}{\sqrt{f^2+(X_c d_x)^2+(Y_c d_y)^2}} \ . \\ x = z \cdot \frac{X_c d_x}{f} \ . &\qquad y = z \cdot \frac{Y_c d_y}{f} \ . \end{aligned} \tag{2}$$

The SR-3000 driver software includes functions performing this transformation from the range map $\mathbf{r}$ to the associated point cloud $\mathcal{P}$, which can be rendered in a 3D visualization software, such as Paraview[4] (fig. 3). This representation illustrates two limitations of range imaging with a TOF camera. First, the lateral field of view (FOV) is limited. In our example, only the top part of the person is included in the image. The second limitation comes from the perspective projection. Objects close to the camera occlude objects farther away. This effect is clearly illustrated by the *shadow* cast by the person over the wall in fig. 3. We will see in the following sections that multi-camera networks allow to overcome those limitations.

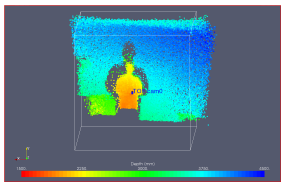


Fig. 3. Cloud of points $\mathcal{P}$ obtained from the TOF range map in figure 2**b**.

1.3 Multi-Camera Operation

Since TOF cameras are active devices with their own illumination source, special care must be taken when multiple cameras are operated simultaneously. In particular, it is desirable to avoid *interference* between different devices, which could lead to erroneous range measurements. In this paper, we will discuss only interference minimization for the SR-3000 camera [1], which provides range data based on the phase measurement of a periodic signal. Büttgen [5] gives a good overview of multi-camera operation options for this device. Options allowing to image the same scene with multiple cameras, without requiring an explicit synchronization of the different devices are Frequency Division Multiple Access (FDMA) and Code Division Multiple Access (CDMA). Although a method using CDMA has been proposed for SwissRanger cameras [5] , its implementation has not yet been released to camera users. Therefore, our experiments were carried out using the FDMA approach. Specifically, we operated a first camera at $f_0 = 20$ MHz and a second camera at $f_1 = 21$ MHz. Lange [6] showed that under these conditions, the crosstalk between the two devices is less than $40\ dB$, as long as the integration time exceeds $T_{int,min} = 100\ \mu s$. Since SR-3000 cameras support 4 different operating frequencies, networks with 4 different SR devices can be used with the FDMA approach. For network with more cameras, the CDMA approach must be employed. Büttgen [5] shows that 15bit long pseudo-noise sequences can be used to define codes allowing to operate up to 1800 cameras without significant interference.

2 Calibration Strategy

In this section, we consider the problem of aligning two point clouds $\mathcal{P}_0$ and $\mathcal{P}_1$ acquired by TOF devices $\mathcal{C}_0$ and $\mathcal{C}_1$, in the specific case where the two fields of view overlap partially. More specifically, the alignment procedure should allow to obtain the rigid body transformation T_{C_0,C_1} between the point of view of cameras $\mathcal{C}_0$ and $\mathcal{C}_1$. This rigid body transformation can be decomposed in a rotation $\mathbf{R}$ and a translation $\mathbf{T}$, and is therefore fully specified by 6 parameters : the three translation coefficients in $\mathbf{T}$, and the rotation angles θ (yaw), ϕ (pitch) and ψ (tilt). If those parameters are known, it becomes possible, for each point $\mathbf{x_1}$ in the cloud $\mathcal{P}_1$, to compute the corresponding position $\mathbf{x_0}$ in the coordinates system of $\mathcal{C}_0$

$$\mathbf{x_0} = T_{C_0,C_1}\left[\mathbf{x_1}\right] = \mathbf{R} \cdot \mathbf{x_1} + \mathbf{T}\ . \tag{3}$$

We do not explicitly consider here the general case of global alignment for a network with more than 2 cameras. Nevertheless, the same approach can be used to align an arbitrary number of cameras, provided that a chain of pairwise overlapping FOV links all cameras in the network. Registration of multiple views of the same scene has been thoroughly investigated for standard (intensity, color) 2D cameras[3][7][8], where very accurate results can be obtained by means of bundle adjustment. We discuss the applicability of this approach for point clouds produced with SR-3000 cameras. Then, we will present an *ad-hoc* registration procedure, based on the range images measured. This method is better suited to the SR-3000 specifications.

2.1 Bundle Adjustment

A classical approach based on a planar target was proposed by Tsai [3] and later refined by Zhang [7] and Bouguet [8], at which point it was integrated into the OpenCV library [9]. In this approach, the planar target is imaged at different positions in the cameras' fields of view. Reference points (usually corners) are extracted in the images. Since the real-world geometry of the target object is known, the perspective projection model for the camera can be inferred from the acquired images by bundle adjustment for the feature points. This photogrammetric calibration procedure provides camera *intrinsic* parameters such as focal length, position of the optical center and distortion, but also *extrinsic* parameters, namely the position of the camera with respect to the target pattern origin. This approach can be used for the SR-3000 camera since an amplitude map $A\,(i,j)$ is measured in addition to the range map $r\,(i,j)$.

As noted above, this method allows to compute extrinsic parameters. If those parameters are matched for images of the *same* target pattern acquired by two devices $\mathcal{C}_0$, $\mathcal{C}_1$ at different positions, the relative position of the two devices can be computed. It is important to note here that this alignment method is based only on intensity images produced by the TOF cameras : the TOF range map is not exploited in this alignment procedure. Kahlmann [10] notes that the sensor lateral resolution is too low to use standard calibration targets for bundle adjustment for the SR-3000 camera. Lindner *et al.* [11] successfully applied the OpenCV calibration procedure to register the range image of a PMD TOF camera with a color image produced by a standard camera (in this case, the lateral resolution of the color imager is higher than the TOF sensor resolution). We carried out experiments based on the calibration toolbox provided by Bouguet [8], using a large checkerboard pattern (each square was $100\,mm$). Figure 4 shows an example of calibration images, along with a rendering of the computed relative camera positions. Section 3.1 illustrates that the extrinsic parameters obtained do not allow to successfully align point clouds obtained with two SR-3000 cameras. One of the key issue is that the camera model parameters computed from the amplitude images may be different from the parameters used in the SR-3000 driver software for the transformation of the range map into a 3D cloud of points. Therefore, we investigated a more robust method to align range images, described in the following section. This method is based on a simple calibration scene, and relies only on the camera model used for point cloud generation in the SR-3000 driver software.

(a) Left image **(b)** Right image **(c)** Computed camera positions

Fig. 4. Bundle adjustment calibration toolbox results

2.2 Surface Matching Techniques

The method presented above is unsatisfactory since it requires the determination of a rigid body transformation between two camera viewpoints. Those viewpoints are usually far away from the measured point clouds, so that any error in the transformation parameters is greatly amplified. Moreover, only a small set of matched points are used to compute the transformation. A significant improvement can be achieved if most of the measured points are considered in the matching procedure. This is the case in automatic surface registration methods such as *iterative closest points* (ICP) [12]. Unfortunately, the TOF data obtained from real-time cameras is very noisy. Construction of valid surface sets from the range map is therefore complicated, since pixel noise generates large surface patches, which would prevent an automatic surface registration procedure from converging. Therefore, we did not further investigate those methods in the present work. Nevertheless, as noise levels and denoising techniques for TOF device will probably improve in the next few years, this solution should be carefully investigated in future works.

We propose here an hybrid method, more suited to the SR-3000 data. We explicitly separate the surface matching in two steps : matching of rotation, and matching of translation between the two point clouds. In order to match the rotation, a planar target is imaged simultaneously with both cameras. Then, a *random sample consensus* (RANSAC) [13] method is used to recognize this planar regions in both point clouds and to compute the normal $\boldsymbol{n}$ to this object in the coordinates system of each camera. This feature is relatively robust to measurement noise, since a large number of points are taken into account for its determination. Point clouds for each camera are subsequently rotated so that this normal becomes parallel to the z axis. Performing the rotation for both cameras ensures that calibration errors are evenly distributed between both clouds. When this step is completed, the remaining translation parameters are computed from a set of 6 to 10 point pairs defined by a human operator. Again, using more points helps in lowering the impact of measurement noise.

Table 1. Intrinsic camera parameters for SR-3100 (sn097027)

Parameter	MESA	calib.
f [mm]	8.0 $\pm$n.a.	8.04 $\pm$0.23
c_x	85.0 $\pm$n.a.	83.5 ± 4.7
c_y	76.7 $\pm$n.a.	80.3 ± 5.4

Table 2. Intrinsic camera parameters for SR-3000 (sn296012)

Parameter	MESA	calib.
f [mm]	8.0 $\pm$n.a.	7.98 $\pm$0.26
c_x	95.1 $\pm$n.a.	93.8 ± 5.2
c_y	56.3 $\pm$n.a.	51.6 ± 5.9

3 Calibration Experiments

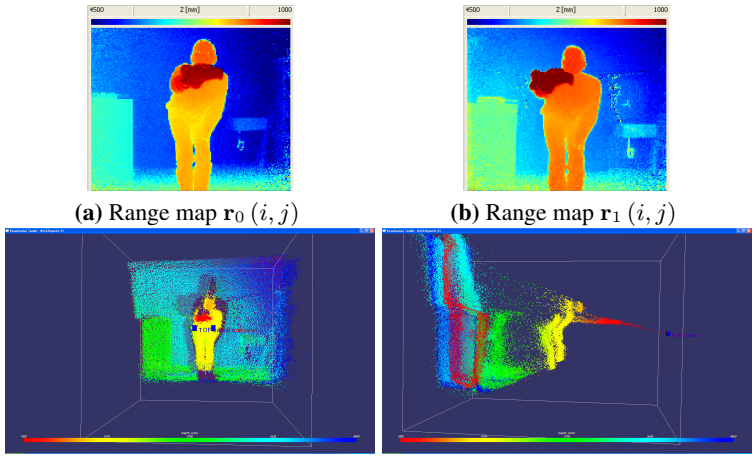
Calibration experiments were carried out using an ad-hoc TOF image acquisition application, allowing to record range map simultaneously with two SR-3000 devices, operated at frequencies $f_0 = 20$ MHz and $f_1 = 21$ MHz respectively. In the experiments presented here, the aim was set to determining the coordinates transformation T_{C_0,C_1} allowing the computation of the position of points $\mathcal{P}_1$ measured by the camera $\mathcal{C}_1$ into the coordinates system of the reference camera $\mathcal{C}_0$. A qualitative evaluation of the calibration performance is therefore possible : point clouds $\mathcal{P}_0$ and $\mathcal{P}_1$ are rendered in the same 3D scene, allowing an human observer to evaluate the fitting of both datasets.

3.1 Camera Calibration Toolbox

This toolbox uses a planar checkerboard pattern as calibration target (fig. 4**a**,**b**), and is based solely on amplitude data. Prior to extrinsic parameters determination, this toolbox allows to determine *intrinsic* parameters of the camera, such as the focal length f or the coordinates (c_x, c_y) of the principal point. In tables 1 and 2 we compare the values obtained with this toolbox to the manufacturer's data for two SR-3X00 devices. In both cases, the agreement is good : the manufacturer's data lies within the uncertainty domains of calibration values. Unfortunately, the *extrinsic* parameters obtained with this calibration method proved useless for point cloud registration (fig. 5). Although disappointing, those bad registration results were expected since the camera model for image registration is (slightly) different from the camera used in point cloud computation. Moreover, variations related to fixed pattern noise at different frequencies further degrade the fitting between the two point clouds.

3.2 RANSAC based Method

To overcome errors due to inaccurate rotation parameters, we used the alignment procedure proposed in section 2.2. For the RANSAC procedure, the number of iterations N was set to 1600 and the threshold for inliers was $d = 650\,mm$. Although this threshold is quite large, the obtained rotation parameters do not suffer greatly from noise,



(a) Range map $\mathbf{r}_0\,(i,j)$ **(b)** Range map $\mathbf{r}_1\,(i,j)$

(e) $\mathcal{P}_0 + \mathcal{P}_1$, view 1 **(f)** $\mathcal{P}_0 + \mathcal{P}_1$, view 2

Fig. 5. Point cloud combination with parameters computed by camera calibration toolbox. Large discrepancies between the point clouds can be observed.

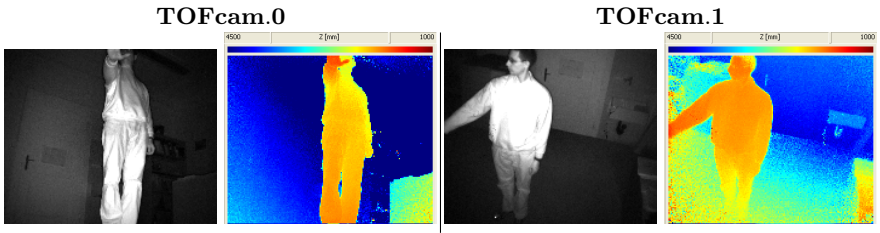
since the plane orientation is determined by averaging on a sufficiently large number of inliers. Typically, good results are obtained when the number of inliers exceeds 5000. The remaining translation degrees of freedom were handled by manually defining 10 point pairs in the amplitude images provided by the SwissRanger cameras. Successful point cloud fusion using this calibration method is illustrated in fig. 8. This example is discussed in the next section.

3.3 Comparison of Calibration Results

The efficiency of the two alignment method discussed above can be compared by employing both methods for the same scene. Table 3 provides a comparison of the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{T}$ for the bundle adjustment method and the RANSAC based method. Large discrepancies can be observed between the two results. Comparison of fused point clouds allows to determine which method performs best. Figure 6 shows an example scene acquired simultaneously with two TOF cameras. As stated above, results obtained with camera calibration toolbox are clearly too inaccurate for successful point cloud merging (fig. 7). On the contrary, the hybrid method where rotation is determined by matching plane primitives obtained through RANSAC allows to successfully fuse the point clouds (fig. 8) : all objects in the scene are uniquely represented in the merged point cloud. In order to go beyond the qualitative discussion presented here, an appropriate distance function between two TOF point clouds must be defined. Work on this topic is currently in progress, and should ultimately allow to determine a least error transformation $(\mathbf{R}, \mathbf{T})$. We provide below preliminary quantitative results for the estimation of the accuracy of the calibration based on RANSAC planes.

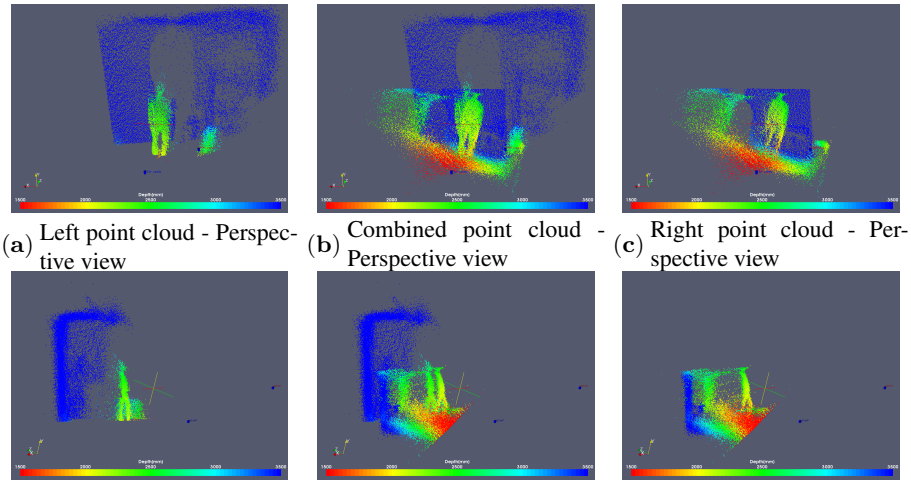
Table 3. Comparison of coordinates transformation determined by different alignment methods

Calibration method	Rotation matrix R	Translation vector T
Bundle adjustment	$\begin{pmatrix} +0.7242 & +0.1296 & +0.6773 \\ +0.3237 & +0.8033 & -0.4999 \\ -0.6088 & +0.5813 & +0.5398 \end{pmatrix}$	$\begin{pmatrix} +1692 \\ -1625 \\ +1292 \end{pmatrix}$
RANSAC plane	$\begin{pmatrix} +0.7404 & +0.1360 & +0.6582 \\ +0.3676 & +0.7380 & -0.5659 \\ -0.5627 & +0.6610 & +0.4964 \end{pmatrix}$	$\begin{pmatrix} +2033 \\ -1957 \\ +140 \end{pmatrix}$



(a) Amplitude **(b)** Range **(c)** Amplitude **(d)** Range

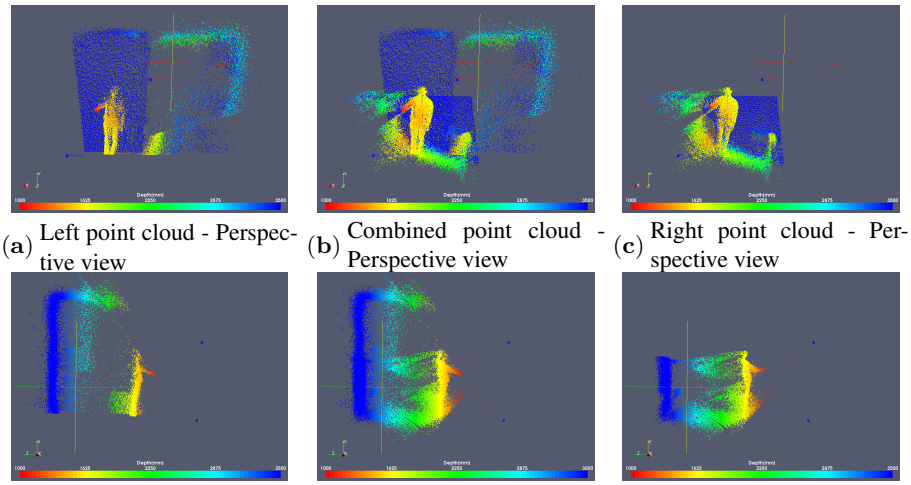
Fig. 6. Amplitude and range images acquired simultaneously with two TOF devices.



(a) Left point cloud - Perspective view (b) Combined point cloud - Perspective view (c) Right point cloud - Perspective view

(d) Left point cloud - Top view (e) Combined point cloud - Top view (f) Right point cloud - Top view

Fig. 7. Point cloud fusion with bundle adjustment calibration. Large discrepancies between the two point clouds can still be observed.



(a) Left point cloud - Perspective view (b) Combined point cloud - Perspective view (c) Right point cloud - Perspective view

(d) Left point cloud - Top view (e) Combined point cloud - Top view (f) Right point cloud - Top view

Fig. 8. Point cloud fusion with RANSAC calibration. The fused point cloud provides more information on the scene than individual clouds.

Accuracy of RANSAC based Calibration Since no ground truth data is available, the accuracy estimation must be carried out by quantifying the correspondence between the 3D point clouds merged. Laveoué et al.[14] proposed a 3D quality metric related to human perception, but this metric is defined only for meshes. Work is currently in progress to define a quality metric adapted to data delivered by TOF camera networks. For preliminary results, we use position measurements of a simple target identified by its amplitude pattern. The target position was measured at 6 different stations. An example set of images for this procedure is given in figure 9. For each station, the distance from the target to the closest camera is larger than 2m. Table 4 summarizes the measured discrepancies for the target position after calibration. The average error is less than 60 mm, for all directions, but the variation is high, indicating that more tests will be

(a) Left camera - Amplitude (b) Right camera - Amplitude (c) Combined point clouds

Fig. 9. Error estimation by measuring the position of a simple target.

Table 4. Discrepancies in position of simple target (6 stations).

Station	**1**	**2**	**3**	**4**	**5**	**6**	**Avg**	**std**
$\mathbf{\Delta x}$ [mm]	+46.3	+70.2	+32.8	+56.7	+49.0	-09.6	**+40.9**	27.6
$\mathbf{\Delta y}$ [mm]	+65.2	+60.3	+61.9	+76.4	+39.0	+32.1	**+55.8**	16.8
$\mathbf{\Delta z}$ [mm]	+13.9	+02.1	+18.1	+20.4	+40.8	+64.4	**+26.6**	22.4

required to confirm this result. Moreover, the observations are not centered around zero, indicating that a systematic error is still present. Nevertheless, the performance achieved is good, especially when compared to the typical noise level for a TOF camera: 150 mm at this range. Those preliminary results confirm that the RANSAC based calibration method performs well in a simple network. Future experiments will help to determine if systematic errors can be further reduced, and if the global accuracy is reduced when more cameras are present in the network.

4 Synchronous Imaging with Calibrated Cameras

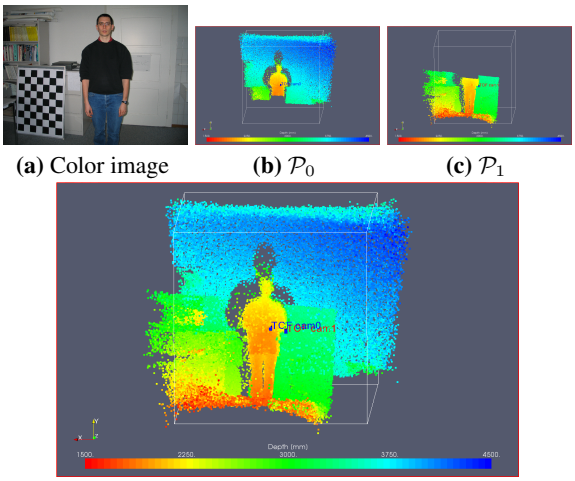
The two problems we are aiming to address with synchronous multi-camera acquisition are extension of the field of view, and removal of occlusion. A TOF camera network allows to perform synchronous acquisition of sequences of range images for scenes containing motion. Parameters for point cloud registration are computed offline, through an interactive procedure using an initial transform provided by the RANSAC registration procedure described in sec. 3.2. The computed point cloud are then visualized in real-time in the same 3D renderer, allowing to produce movies of 3D scenes.

4.1 FOV Extension

Figure 10 illustrates field of view extension. In this experiment, a person walks in front of two TOF cameras. The reference camera $\mathcal{C}_0$ is aimed at the upper body, while the second camera image toward the legs. The resulting point clouds clearly shows the reconstruction of the full person walking. We emphasize here that the fusion is a *real-time* operation. The example of figure 10 is just a frozen frame of a sequence recorded at 20 fps.

4.2 Occlusion Removal

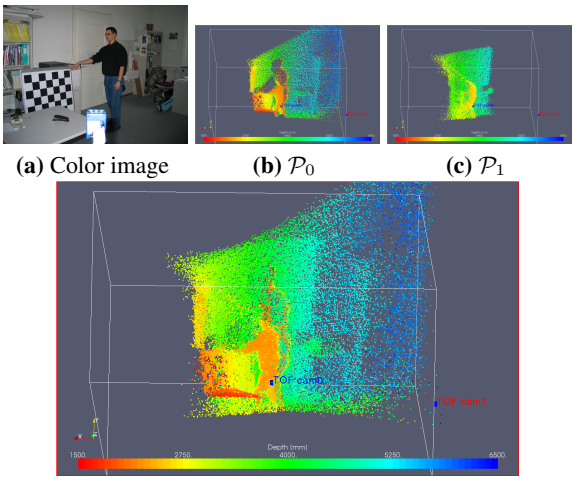
Removal of occlusions is illustrated in fig. 11. Using a single TOF camera results in large unmonitored areas on the wall behind a person close to the camera (fig. 11**b**). The occlusion is removed when a second TOF camera is added (fig. 11**d**), allowing to confirm that a single human is present in the cameras field of view. A typical application of occlusion resolution is human monitoring applications, where the number of humans present in a defined area must be reliably determined.



(a) Color image **(b)** $\mathcal{P}_0$ **(c)** $\mathcal{P}_1$

(d) $\mathcal{P}_0 + \mathcal{P}_1$

Fig. 10. Point clouds combination for field of view extension.



(a) Color image **(b)** $\mathcal{P}_0$ **(c)** $\mathcal{P}_1$

(d) $\mathcal{P}_0 + \mathcal{P}_1$

Fig. 11. Point clouds combination for occlusions removal.

5 Conclusions

Applications involving synchronous operation TOF cameras in a multi-camera network were investigated. Experiments carried out with SR-3000 cameras show that registration methods used for conventional cameras can not be applied directly to TOF imagers with coarse sampling. An hybrid method using RANSAC plane primitives for alignment was introduced. This method provides better results, and allows to fuse points clouds acquired by the TOF network. While calibration is performed offline, fusion is carried out in real-time, allowing the network to operate at the TOF camera native speed (20 fps). Field of view extension and occlusion removal were realized with multi-camera networks, for scenes with motion. Possible improvements to the present work include definition of an error metric for point cloud fusion allowing to determine a least error solution, and extensions to support this error minimization for an arbitrary number of TOF sensors in the network.

References

1. Mesa Imaging: Swissranger SR-3000 (2007) http://www.mesa-imaging.ch/prodviews.php.
2. Canesta Inc.: CanestaVision (2006) http://www.canesta.com/index.htm.
3. Tsai, R.Y.: A versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses. IEEE Journal of Robotics and Automation **RA-3** (1987) 323–344
4. Kitware: Paraview - Parallel visualization application (2007) http://www.paraview.org.
5. Büttgen, B.: Extending time-of-flight optical 3D-imaging to extreme operating conditions. PhD dissertation, Université de Neuchâtel (2007)
6. Lange, R.: 3D Time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. PhD dissertation, University of Siegen (2000)
7. Zhang, Z.: Flexible Camera Calibration By Viewing a Plane From Unknown Orientations. In: International Conference on Computer Vision (ICCV'99), IEEE (1999) 666–673
8. Bouguet, J.Y.: Camera calibration toolbox for matlab (2004) http://www.vision.caltech.edu/bouguetj/calib_doc/ (accessed on 10.09.2007).
9. Intel Corp.: OpenCV (2008) http://www.intel.com/technology/computing/opencv/ (accessed on 15.02.2008).
10. Kahlmann, T.: Range imaging metrology : investigation, calibration and development. PhD dissertation, Eidgenössische Technische Hochschule ETH Zürich (2007) Diss. ETH 17392.
11. Lindner, M., Kolb, A.: Data Fusion and Edge-Enhanced Distance Refinement for 2D RGB and 3D Range Images. In: Proc. of the Int. IEEE Symp. on Signals, Circuits & Systems (ISSCS). Volume 1. (2007) 121–124
12. Jost, T., Hügli, H.: A Multi-Resolution ICP with Heuristic Closest Point Search for Fast and Robust 3D Registration of Range Images. In: Proc. 3Dim2003, 4th Int. Conf. on 3-D Digital Imaging and Modeling. Volume PR01991. (2004) 427–433
13. Torr, P., Murray, D.: The development and comparison of robust methods for estimating the fundamental matrix. International Journal of Computer Vision **24** (1997) 271–300
14. Lavoué, G., Drelie Gelasca, E., Dupont, F., Baskourt, A., Ebrahimi, T.: Perceptually driven 3D distance metrics with application to watermarking. In: Proc. SPIE 6312. (2006)