Photoconsistent Relative Pose Estimation Between a PMD 2D3D-Camera and Multiple Intensity Cameras

Christian Beder, Ingo Schiller, and Reinhard Koch

Computer Science Department
Kiel University, Germany
{beder,ischiller,rk}@mip.informatik.uni-kiel.de

Abstract. Active range cameras based on the Photonic Mixer Device (PMD) allow to capture low-resolution depth images of dynamic scenes at high frame rates. To use such devices together with high resolution optical cameras (e.g. in media production) the relative pose of the cameras with respect to each other has to be determined. This task becomes even more challenging, if the camera is to be moved and the scene is highly dynamic.
We will present an efficient algorithm for the estimation of the relative pose between a single 2D3D-camera with respect to several optical cameras. The camera geometry together with an intensity consistency criterion will be used to derive a suitable cost function, which will be optimized using gradient descend. It will be shown, how the gradient of the cost function can be efficiently computed from the gradient images of the high resolution optical cameras.
We will show, that the proposed method allows to track and to refine the pose of a moving 2D3D-camera for fully dynamic scenes.

1 Introduction

In recent years active range cameras based on the Photonic Mixer Device (PMD) have become available. Those cameras deliver low resolution depth images at high frame rates comparable to usual video cameras. Some PMD-cameras also capture intensity images registered with the depth images (c.f. [11] or [10]) and are therefore sometimes called 2D3D-cameras.

In some applications, such as media production, the images of the 2D3D-camera cannot be used directly. In those cases high resolution optical cameras are used together with the depth information from 2D3D-cameras. The combination of PMD and optical cameras has for instance been used in [6], which explicitly does not require an accurate relative orientation between the two systems.

A combination of PMD and stereo images has also been proposed in [9] and in [1], which rely on an accurate relative orientation between the optical cameras and the PMD-camera. There, the PMD-camera is not allowed to be moved and a fixed rig is used instead, which is calibrated beforehand using a calibration pattern (cf. [2] and [13]).

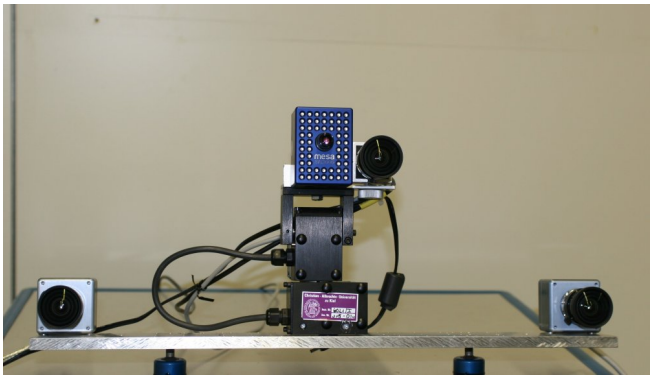

Fig. 1. The setup comprising of a moving PMD 2D3D-camera mounted on a pan-tilt unit between two fixed high resolution optical cameras.

Due to the narrow opening angle and the low resolution of the PMD camera a fixed rig limits the visible space of the 2D3D-camera. This limitation can be circumvented by moving the 2D3D-camera and focus on the interesting spots in the scene. See figure 1 for the setup we used comprising of two high resolution optical cameras and a 2D3D-camera mounted on a pan-tilt unit.

However, in case the 2D3D-camera is moving, its relative pose with respect to the optical cameras has to be determined. Tracking the pose of a moving PMD-camera has been presented in [3] and also in [5]. There the intensity information from the 2D3D-camera is not used. Tracking the pose of a moving 2D3D-camera using also the intensity information has been done in [8], [12] and [14]. In contrast to our approach all those approaches are not based on additional optical cameras and therefore require a static scene. Because we estimate the relative pose between the 2D3D-camera and the optical cameras for each frame, our algorithm is also able to cope with fully dynamic scenes as long as the images are synchronized.

This paper is organized as follows: in section 2 we will derive the geometry of the 2D3D-camera in relation to the optical cameras and propose a cost function based on an intensity consistency constraint. In section 3 it will be shown, how this cost function can be efficiently optimized using gradient descend. Finally we will present some results on synthetic and real data taken with the setup depicted in figure 1 in section 4.

2 Camera Geometry and Intensity Consistency

First we will introduce some notation describing the geometry of the cameras. We start with the 2D3D-camera, from which we obtain intensity as well as depth images

$$I_0[\boldsymbol{x}] : I\!R^2 \mapsto I\!R \qquad D_0[\boldsymbol{x}] : I\!R^2 \mapsto I\!R \tag{1}$$

for each shot. The camera geometry of this 2D3D-camera will be characterized by the projection matrix (cf. [7, p.143])

$$\mathsf{P}_0 = \mathsf{K}_0 R_0^\mathsf{T}(I_3| - \boldsymbol{C}_0) \tag{2}$$

which maps homogeneous 3d points $\mathbf{X}$ to homogeneous image coordinates according to

$$\mathbf{x} = \mathsf{P}_0 \mathbf{X} \tag{3}$$

In our application the cameras are assumed to be calibrated, i.e. we assume K_0 to be known. The pose of the 2D3D-camera, i.e. R_0 and $\boldsymbol{C}_0$, is unknown and will be estimated in the following.

Inverting equation (3) yields for each pixel $\boldsymbol{x}$ together with the depth image $D_0[\boldsymbol{x}]$ a 3d point

$$\boldsymbol{X} = \frac{D_0[\boldsymbol{x}] R_0 K_0^{-1} \mathbf{x}}{\sqrt{\mathbf{x}^\mathsf{T} \mathsf{K}_0^{-\mathsf{T}} \mathsf{K}_0^{-1} \mathbf{x}}} + \boldsymbol{C}_0 \tag{4}$$

We now introduce additional intensity cameras, which produce normal intensity images

$$I_i[\boldsymbol{x}_i] : I\!R^2 \mapsto I\!R \tag{5}$$

triggered synchronously with the 2D3D-camera. The camera geometry of those additional cameras will be described by the projection matrices

$$\mathsf{P}_i = \mathsf{K}_i R_i^\mathsf{T}(I_3| - \boldsymbol{C}_i) \tag{6}$$

which are assumed to be completely known in the following. The 3d points from equation (4) are then projected into those intensity cameras at the homogeneous coordinates

$$\mathbf{x}_i = \mathsf{K}_i R_i^\mathsf{T}(\boldsymbol{X} - \boldsymbol{C}_i) \tag{7}$$

Denoting the homogeneous image coordinates as

$$\mathbf{x}_i = \begin{pmatrix} u_i \\ v_i \\ w_i \end{pmatrix} \tag{8}$$

the Euclidean coordinates are obtained by simple normalization

$$\boldsymbol{x}_i = \frac{1}{w_i} \begin{pmatrix} u_i \\ v_i \end{pmatrix} \tag{9}$$

We have now established correspondences between pixels $\boldsymbol{x}$ in the 2D3D-camera with a pixel in each intensity camera $\boldsymbol{x}_i$ using the depth image $D[\boldsymbol{x}]$. Using the derived pixel correspondences $\boldsymbol{x} \leftrightarrow \boldsymbol{x}_i$, we will assume that the intensities of those pixels are equal up a per image brightness offset b_i and a per image contrast difference c_i

$$c_i I_0[\boldsymbol{x}] \stackrel{!}{=} I_i[\boldsymbol{x}_i] + b_i \tag{10}$$

This condition only holds true, if the pixel is not occluded. Therefore we introduce an occlusion map on the images of the 2D3D-camera

$$\nu_i[\boldsymbol{x}] = \begin{cases} 1 \text{ if } \boldsymbol{x} \leftrightarrow \boldsymbol{x}_i \text{ is not occluded} \\ 0 \qquad\qquad\qquad\quad \text{otherwise} \end{cases} \tag{11}$$

which can be computed from the depth map D using shadow mapping techniques (cf. [15]). Note, that approximate pose parameters R_0 and $\boldsymbol{C}_0$ are required for this operation.

Finally introducing the robust cost function

$$\Psi[x] = \begin{cases} x^2 & \text{if } |x| < \theta \\ |x| + \theta^2 - \theta & \text{otherwise} \end{cases} \tag{12}$$

we are able to formulate the intensity consistency constraint by optimizing the following cost function

$$\phi(\mathsf{R}_0, \boldsymbol{C}_0, \boldsymbol{c}, \boldsymbol{b}) = \sum_i \sum_{\boldsymbol{x}} \nu_i[\boldsymbol{x}] \Psi[I_i[\boldsymbol{x}_i] - c_i I_0[\boldsymbol{x}] + b_i] \rightarrow \min \tag{13}$$

for the unknown pose parameters R_0 and $\boldsymbol{C}_0$ as well as the unknown brightness offsets $\boldsymbol{b}$ and contrast differences $\boldsymbol{c}$. In case of a PMD 2D3D-camera the resolution and hence the number of pixels is small, so that we sum over all pixels $\boldsymbol{x}$ of the image. To improve the running time of the algorithm in case of higher resolution 2D3D-cameras it is also possible to detect interest points and sum up only those.

In the following section we will show, how this cost function can be efficiently optimized using gradient descent techniques.

3 Optimization

We will now show, how the cost function derived in the previous section can be efficiently optimized. Therefore we first approximate the rotation matrix by (cf. [4, p.53])

$$\mathsf{R}_0 = \overline{\mathsf{R}}_0 + [\boldsymbol{r}_0]_\times \tag{14}$$

where $[\cdot]_\times$ denotes the 3×3 skew symmetric matrix induced by the cross product (cf. [7, p.546]). The unknown rotation is now parameterized using the 3-vector $\boldsymbol{r}_0$ containing the differential rotation angles. Stacking all the unknown parameters into the parameter vector

$$\boldsymbol{p} = \begin{pmatrix} \boldsymbol{r}_0 \\ \boldsymbol{C}_0 \\ \boldsymbol{c} \\ \boldsymbol{b} \end{pmatrix} \tag{15}$$

the gradient of the cost function (13) is given by

$$\boldsymbol{g} = \sum_i \sum_{\boldsymbol{x}} \nu_i[\boldsymbol{x}] \Psi'[I_i[\boldsymbol{x}_i] - c_i I_0[\boldsymbol{x}] + b_i] \left(\frac{\partial I_i[\boldsymbol{x}_i]}{\partial \boldsymbol{p}} - I_0[\boldsymbol{x}] \boldsymbol{e}_{i+6}^\mathsf{T} + \boldsymbol{e}_{i+N+6}^\mathsf{T} \right) \tag{16}$$

where N is the number of images and $\boldsymbol{e}_i$ is a vector of the same size as the parameter vector $\boldsymbol{p}$ containing zeros except for the i-th component, which is one. Note, that the dependence of the occlusion map $\nu_i[\boldsymbol{x}]$ on the rotation and translation of the 2D3D-camera is neglected here.

We will now look at the components of the gradient. First, the derivative of the robust cost function is simply given by

$$\Psi'[x] = \begin{cases} 2x & \text{if } |x| < \theta \\ 1 & \text{otherwise} \end{cases} \tag{17}$$

Second, the partial derivatives of the intensity images $I_i[\boldsymbol{x}_i]$ with respect to the unknown parameters $\boldsymbol{p}$ is obtained from the gradient images $\nabla I_i[\boldsymbol{x}_i]$ using chain rule as

$$\frac{\partial I_i[\boldsymbol{x}_i]}{\partial \boldsymbol{p}} = \nabla I_i[\boldsymbol{x}_i] \frac{\partial \boldsymbol{x}_i}{\partial \mathbf{x}_i} \frac{\partial \mathbf{x}_i}{\partial \boldsymbol{X}} \frac{\partial \boldsymbol{X}}{\partial \boldsymbol{p}} \tag{18}$$

Its components are the partial derivatives of equation (7) being

$$\frac{\partial \boldsymbol{x}_i}{\partial \mathbf{x}_i} = \begin{pmatrix} \frac{1}{w_i} & 0 & -\frac{u_i}{w_i^2} \\ 0 & \frac{1}{w_i} & -\frac{v_i}{w_i^2} \end{pmatrix} \tag{19}$$

as well as of equation (9) being

$$\frac{\partial \mathbf{x}_i}{\partial \boldsymbol{X}} = \mathsf{K}_i \mathsf{R}_i^\mathsf{T} \tag{20}$$

Finally we need the Jacobian

$$\frac{\partial \boldsymbol{X}}{\partial \boldsymbol{p}} = \begin{pmatrix} \frac{\partial \boldsymbol{X}}{\partial \boldsymbol{r}_0} & \frac{\partial \boldsymbol{X}}{\partial \boldsymbol{C}_0} & \mathsf{0}_{3\times 2N} \end{pmatrix} \tag{21}$$

where N is the number of intensity images and the null matrix $\mathsf{0}_{3\times 2N}$ reflects the fact, that the 3d point is independent of the image brightness and contrast. The components of this remaining Jacobian are the partial derivatives of equation (4) given by

$$\frac{\partial \boldsymbol{X}}{\partial \boldsymbol{r}_0} = \frac{-D_0[\boldsymbol{x}]}{\sqrt{\mathbf{x}^\mathsf{T} \mathsf{K}_0^{-\mathsf{T}} \mathsf{K}_0^{-1} \mathbf{x}}} \left[\mathsf{K}_0^{-1} \mathbf{x}\right]_\times \tag{22}$$

as well as

$$\frac{\partial \boldsymbol{X}}{\partial \boldsymbol{C}_0} = \mathsf{I}_3 \tag{23}$$

Note, that only the gradient images of the high resolution optical cameras and some simple matrix operations are required to compute the gradient, which can be performed very efficiently.

Having derived an efficient method for computing the gradient of the cost function from the gradient intensity images, it is now possible to minimize the cost function using gradient descent techniques. To do so we need to determine a step width. This can be done by using the curvature of the cost function in

the direction of the gradient. We therefore compute three samples of the cost function

$$\phi_0 = \phi(\boldsymbol{p}_0) \qquad \phi_1 = \phi(\boldsymbol{p}_0 - \epsilon \boldsymbol{g}) \qquad \phi_2 = \phi(\boldsymbol{p}_0 - 2\epsilon \boldsymbol{g}) \tag{24}$$

and note, that the parabola fitted through this three values has its peak at

$$\epsilon_0 = \epsilon \frac{3\phi_0 - 4\phi_1 + \phi_2}{2\phi_0 - 4\phi_1 + 2\phi_2} \tag{25}$$

If the cost function in the direction of the gradient is convex, then $\epsilon_0 > 0$. If this is not the case, we set ϵ_0 to an arbitrary positive value (e.g. $\epsilon_0 = \epsilon$). Then we proceed as follows: Starting with a line-search factor $\alpha = 1$ we check, if there is an improvement in the cost function, i.e. if

$$\phi(\boldsymbol{p}_0 - \alpha\epsilon_0 \boldsymbol{g}) < \phi(\boldsymbol{p}_0) \tag{26}$$

If this is the case, we update $\boldsymbol{p}_0$ accordingly and iterate. If this is not the case we reduce α and repeat the test. This scheme is iterated, until the parameter updates are sufficiently small.

We have presented an efficient estimation scheme, which we will evaluate in the following section.

4 Results

To evaluate the proposed algorithm, we rendered a synthetic data set from a known geometry, for which ground truth data is available. The setup was chosen similar to the one depicted in figure 1 comprising of two optical cameras with a 2D3D-camera in between. Figure 2 shows the 3d setup we used. The optical cameras are placed $1m$ to the left and to the right of the 2D3D-camera at a distance of about $18m$ from the castle with the occluding arc at a distance of approximately $11m$. The resolution of the optical cameras was chosen as 640×480 pixels at an opening angle of $70°$ while the resolution of the 2D3D-camera was chosen as 160×120 pixels at an opening angle of $40°$.

On the left hand side of figure 3 the image from one of the optical cameras is shown together with the image of the 2D3D-camera re-projected using exemplary erroneous initial pose parameters at a distance of $0.4m$ from the ground truth. On the right hand side of figure 3 the same image is shown after the optimization. As expected the registration between the images has improved.

In order to quantify this improvement, we initialized the algorithm with pose parameters disturbed by Gaussian noise of increasing standard deviation σ_C and compared the resulting poses with the known ground truth poses. In addition to this disturbance of the initial pose parameters we repeated the experiment with Gaussian noise on the depth images of the 2D3D-camera with standard deviation $\sigma_{depth} = 2m$ as well as with Gaussian noise on the optical images with standard deviation $\sigma_{int} = 50$ at an intensity range of 255. The distances d_c of the resulting poses to the known ground truth poses together with their standard deviations are plotted against the standard deviation of the disturbance of the initial pose

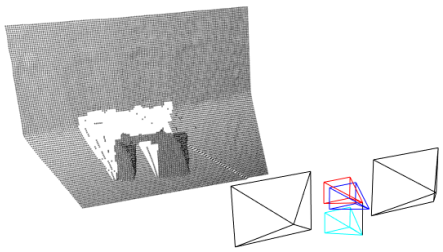

Fig. 2. Estimated pose and point cloud after optimization of the synthetic scene. The camera symbols in the middle show the pose of the 2D3D-camera before the optimization, after the first iteration and at the final position.

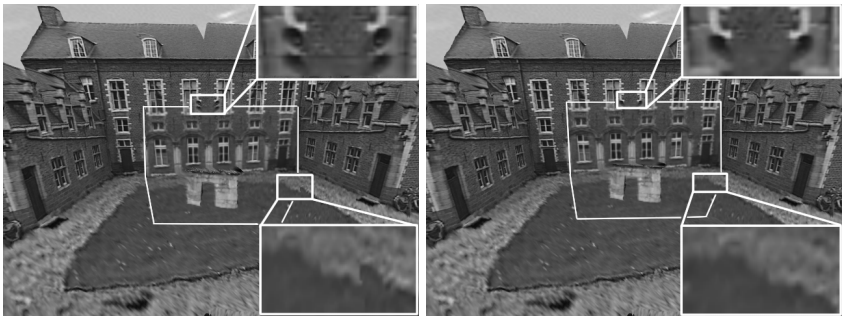

Fig. 3. Left image of the synthetic sequence. The 2D3D image is overlaid onto the intensity image using the initial pose on the left hand side and using the optimized pose on the right hand side. The enlargements show that the registration error decreases. Note the speakers in the top view and the corrected discontinuity on the bottom view.

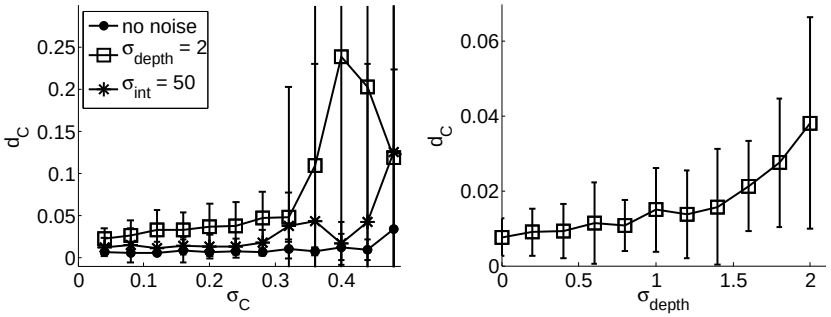

Fig. 4. *Left:* Mean distance and standard deviation of the resulting pose to the ground truth pose plotted against the standard deviation of the initial pose disturbance. For the bottom curve no noise was added to the input images, while the two other curves result from severe additional noise in the depth and intensity images respectively. *Right:* Mean distance and standard deviation of the resulting pose to the ground truth pose plotted against the standard deviation of the depth image noise using initial pose disturbance with standard deviation $\sigma_C = 0.25m$.

parameters σ_C on the left hand side of figure 4. Observe, that the algorithm reacts much more sensitive to disturbances of the depth than to disturbances of the intensity images. The sensitivity of the algorithm to disturbances of the depth images is shown on the right hand side of figure 4, where we plotted the distances d_c of the resulting poses to the known ground truth poses together with their standard deviations against the standard deviation of the disturbance of the depth images σ_{depth} while keeping the disturbance of the initial pose at $\sigma_C = 0.25m$. It can be seen, that in this example severe disturbances of depth and intensity images do not affect the performance of the algorithm unless the pose initialization is above $\sigma_C = 0.25m$. As expected the Gaussian intensity noise is well compensated by the robustified least-squares optimization. It can be seen, that also the Gaussian noise in the depth measurements, which also affects the occlusion maps, can be coped with to some degree.

Fig. 5. Left image of a real sequence. The 2D3D image is overlaid onto the intensity image using the initial approximate pose from the pan-tilt unit on the left hand side and using the optimized pose on the right hand side. Note the corrected registration for instance on the enlarged part of the poster in the background.

Next, we evaluated our algorithm on real images taken with the setup depicted in figure 1. It comprises of two optical cameras with a resolution of 1600×1200 pixels having an opening angle of $60°$ in combination with a 2D3D-camera mounted on a pan-tilt unit. The 2D3D-camera comprises of a PMD camera with a resolution of 176×144 pixels having an opening angle of $40°$ operated together with a rigidly coupled and calibrated optical camera from which the intensity values are taken using the mutual calibration. Figure 5 shows the intensity image of the left camera together with the re-projection of the image of the 2D3D-camera. On the left hand side the initial pose obtained approximately from the pan-tilt unit is used and the image on the right hand side shows the re-projection after the optimization. Observe, that on the left hand side the registration is fairly poor, while the mutual registration after the optimization is significantly improved as shown in the image on the right hand side. The 3d camera poses before and after the optimization are depicted in figure 6.

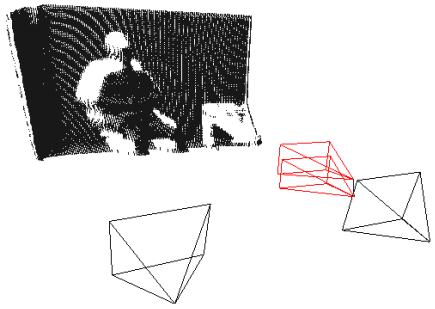

Fig. 6. Estimated pose and point cloud after optimization of the real scene. Again the camera symbols in the middle show the pose of the 2D3D-camera before and after optimization.

5 Conclusion

We have presented a system for estimating the pose of a moving 2D3D-camera with respect to several high resolution optical cameras. Because the method is based on single synchronized shots of the scene, no additional assumptions on the rigidness are made and the scene is allowed to be fully dynamic.

It has been shown, that our approach is quite robust and able to cope very well with severe noise on the intensity as well as on the depth images. The presented method is a greedy gradient based optimization, so that we require good initial values for the pose parameters. Those can be either obtained from external sources, such as the rotation data from the pan-tilt unit or an inertial sensor mounted on the 2D3D-camera, or by tracking the pose over an image sequence. Our experiments indicate, that both approaches are feasible and the radius of convergence is sufficiently large for the application of the pan-tilt unit or the inertial sensor as well as for tracking the pose at high frame rates.

As the radius of convergence is governed by the reach of the image gradients, future work will focus on improving the convergence by introducing a multi-scale coarse-to-fine optimization. We expect, that thereby the tracking of faster rotational movements, which cause large image displacements even at high frame rates, can be improved.

Acknowledgments

This work was partially supported by the German Research Foundation (DFG), KO-2044/3-1, and the Project 3D4YOU, Grant 215075 of the Information Society Technologies area of the EUs 7^{th} Framework programme.

References

1. Christian Beder, Bogumil Bartczak, and Reinhard Koch. A combined approach for estimating patchlets from PMD depth images and stereo intensity images. In F.A. Hamprecht, C. Schnörr, and B. Jähne, editors, *Proceedings of the DAGM 2007*, number 4713 in LNCS, pages 11–20. Springer, 2007.
2. Christian Beder and Reinhard Koch. Calibration of focal length and 3d pose based on the reflectance and depth image of a planar object. In *Proceedings of the DAGM Dyn3D Workshop, Heidelberg, Germany*, 2007.
3. Christian Beder and Reinhard Koch. Real-time estimation of the camera path from a sequence of intrinsically calibrated pmd depth images. In *Proceedings of the ISPRS Congress, Bejing, China*, 2008. to appear.
4. Wolfgang Förstner and Bernhard Wrobel. Mathematical concepts in photogrammetry. In J.C.McGlone, E.M.Mikhail, and J.Bethel, editors, *Manual of Photogrammetry, Fifth Edition*, pages 15–180. ASPRS, 2004.
5. Stefan Fuchs and Stefan May. Calibration and registration for precise surface reconstruction with tof cameras. In *Proceedings of the DAGM Dyn3D Workshop, Heidelberg, Germany*, 2007.
6. Uwe Hahne and Marc Alexa. Combining time-of-flight depth and stereo images without accurate extrinsic calibration. In *Proceedings of the DAGM Dyn3D Workshop, Heidelberg, Germany*, 2007.
7. R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.
8. Benjamin Huhle, Philipp Jenke, and Wolfgang Strasser. On-the-fly scene acquisition with a handy multisensor-system. In *Proceedings of the DAGM Dyn3D Workshop, Heidelberg, Germany*, 2007.
9. K.D. Kuhnert and M. Stommel. Fusion of stereo-camera and PMD-camera data for real-time suited precise 3d environment reconstruction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2006.
10. Marvin Lindner, Martin Lambers, and Andreas Kolb. Sub-pixel data fusion and edge-enhanced distance refinement for 2d/3d images. In *Proceedings of the DAGM Dyn3D Workshop, Heidelberg, Germany*, 2007.
11. T.D.A. Prasad, K. Hartmann, W. Wolfgang, S.E. Ghobadi, and A. Sluiter. First steps in enhancing 3d vision technique using 2d/3d sensors. In V. Chum, O.Franc, editor, *Computer Vision Winter Workshop 2006*, pages 82–86, University of Siegen, 2006. Czech Society for Cybernetics and Informatics.
12. A. Prusak, O. Melnychuk, Ingo Schiller, H. Roth, and R. Koch. Pose estimation and map building with a pmd-camera for robot navigation. In *Proceedings of the DAGM Dyn3D Workshop, Heidelberg, Germany*, 2007.
13. Ingo Schiller, Christian Beder, and Reinhard Koch. Calibration of a pmd camera using a planar calibration object together with a multi-camera setup. In *Proceedings of the ISPRS Congress, Bejing, China*, 2008. to appear.
14. B Streckel, B. Bartczak, R. Koch, and A. Kolb. Supporting structure from motion with a 3d-range-camera. In *Scandinavian Conference on Image Analysis (SCIA07)*, June 2007.
15. Lance Williams. Casting curved shadows on curved surfaces. *SIGGRAPH Comput. Graph.*, 12(3):270–274, 1978.