Supporting Structure from Motion with a 3D-Range-Camera

Birger Streckel[1], Bogumil Bartczak[1], Reinhard Koch[1], and Andreas Kolb[2]

[1] Institute of Computer Science
Christian-Albrechts-University of Kiel, 24098 Kiel, Germany
rk@informatik.uni-kiel.de
[2] Computer Graphics Group
University of Siegen, 57068 Siegen, Germany

Abstract. Tracking of a camera pose in all 6 degrees of freedom is a task with many applications in 3D-imaging as i.e. augmentation or robot navigation. Structure from motion is a well known approach for this task, with several well known restrictions. These are namely the scale ambiguity of the calculated relative pose and the need of a certain camera movement (preferably lateral) to initiate the tracking.

In the last few years time-of-flight imaging sensors were developed that allow the measuring of metric depth over a whole region with a frame rate similar to a standard CCD-camera.

In this work a camera rig consisting of a standard 2D CCD camera and a time-of-flight 3D camera is used. Structure from motion is calculated on the 2D image, aided by the depth measurement from the time-of-flight camera to overcome the restrictions named above. It is shown how the additional 3D-information can be used to improve the accuracy of the camera pose estimation.

1 Introduction

Determining the position of a single moving video camera without the help of markers is a well researched problem. One of the most promising solutions is the Structure From Motion (SfM) technique, where simultaneous to the camera pose estimation a sparse scene structure is reconstructed from prominent 2D-features, that are tracked throughout a video sequence.

A recent approach to measure the shape of a surface in a field of view (FOV) is the PMD-Camera. This camera is using the Photo Mixing Detector (PMD) technology to measure depth values in a FOV, with a frame rate comparable to a common CCD video camera. The PMD-Camera is based on the time-of-flight principle [9]. This paper is discussing how a time-of-flight camera can be used to aid SfM and which benefits and limitations can be expected.

2 Previous Work

Much work was undertaken in the field of camera tracking and SLAM [2] using a single moving 2D-camera. The main contributions were gathered by Hartley and

B.K. Ersbøll and K.S. Pedersen (Eds.): SCIA 2007, LNCS 4522, pp. 233–242, 2007.

Zissermann in [3], a profound overview was given by Lepetit and Fua in [10]. There was also much research done in the field of 3D-imaging based on the time-of-flight principle [15][7].

Using the depth images of a time-of-flight camera to improve camera tracking in 6 degrees of freedom is a fairly new idea and is based on the work of Prasad et al. [12]. There a low resolution 3D-imaging sensor is combined with a conventional high resolution 2D-CCD. The pixel correspondence is assured by an optical image multiplier which is delivering the same optical information to both measuring devices. The high resolution image information of this device enables SfM calculation aided by a known corresponding depth.

2.1 Structure from Motion

The basic idea of SfM is discussed in [11], where a single camera is moved in a static scene and simultaneously the camera position is tracked and a sparse 3D-model of the scene is generated. This method works in realtime [4] without any prior knowledge of the scene, especially without knowing the scene depth.

There are approaches where already depth information is used to aid the SfM task. Koeser et al. [5] i.e. create a scene model in an offline phase. This model is used to generate 3D-points in the online phase, where the camera pose can be estimated in realtime.

There are restrictions in the SfM approach, that limit its usability for certain applications as i.e. robot navigation. One restriction is the necessity of an initial camera movement to be able to start the tracking. A lateral movement leads to better initialization results than a forward movement. In addition the camera pose and scene structure is computed "up to scale", which means that all gathered distances are scaled with an arbitrary factor.

Based on only SfM navigation a robot would have to start moving without knowing anything about its surrounding and also would be unable to detect the real metric distance to an obstacle in its way. In addition the direction for a common robot movement is forward. This is an ill posed problem for the standard SfM approach.

2.2 Photo Mixing Detector (PMD) Technology

The time-of-flight technique measures the metric distance to an object. Modulated light is sent out, reflected by the object and received by an appropriate detector. By sampling and correlating the incoming optical signal with a reference signal it is possible to calculate the time-of-flight for the light ray. Knowing the time, it is easy to calculate the distance the light covered from sending to receiving and thus the distance of the reflecting object.

The PMD technology uses the time-of-flight principle to measure the depth over a whole FOV [15]. The scene is illuminated by LEDs sending out modulated near infrared light. The light is reflected by all visible objects and received by an image sensor, that is comparable to a CMOS chip of a common digital camera.

Also the manufacturing corresponds to the standard CMOS manufacturing process. This allows a very economic production of the device. Due to an automatic suppression of background light the sensor can be used for indoor as well as for outdoor scenes, which is a strong precondition for mobile robot navigation.

Current devices provide a resolution of 64×48 with active background light suppression, the maximum FOV is 40°. The restriction in the FOV is due to the necessity of a bright active illumination of the measured scene. The camera frame rate is 25 Hz and the modulation frequency is 20 MHz, resulting in an unambiguous range of $7.5m$.

2.3 Camera Setup and Test Sequences

The performance increase of SfM with 3D-imaging was measured on simulated image sequences as well as on real camera data.

The real image sequences were generated using a rig of a high resolution 2D camera rigidly fixed to a low resolution PMD-camera, figure 1 shows the setup. We used a PMD-camera from PMD-Tech [7], based on the technology described in section 2.2, with a resolution of 64×48 and a horizontal FOV of only 23.1°. The 2D camera was a standard CCD-camera with a resolution of 1024×768 and a horizontal FOV of 42.3°. The alignment of the cameras assured that the FOV of the PDM-camera was completely covered by the 2D camera. For calibrating the rig we used the Bouguet calibration toolbox [1] similar to Kolb et al. in [6]. With this calibration the depth data was mapped to the 2D-image, an image pair is shown in figure 1. In recent publications the accuracy of the PMD-camera was evaluated [8] and the measured depth values were calibrated [6]. The results from these publications were used to get well calibrated depth data with known uncertainty.

For the real image sequence no ground truth information is available, so it is necessary to evaluate the pose tracking performance also on synthetic data, where we are able to quantify the estimation results. The simulated 2D-images have a resolution of 1024×768 and a FOV of 80°, the depth images have a resolution of 64×48 pixel and a FOV of 40°. Image pairs from the simulated sequence are shown in figure 2. We use the results of [8] to simulate depth noise for the synthetic PMD-images.

Fig. 1. 2D/3D-camera rig with image pair, PMD-image mapped to match 2D-image

3 Using Depth Information for Camera Tracking

The usage of metric depth images for camera pose estimation has three main advantages:

- The camera poses can be estimated in a metric coordinate system.
- The camera pose estimation starts from the first frame, no initial movement is necessary (see section 3.1).
- The camera pose estimation is improved by using the additional depth information.

How these improvements are achieved is described in the following sections.

3.1 Metric Pose Estimation from the First Image

Traditional SfM camera tracking always needs two images for initialization. These images must provide an adequate baseline in order to triangulate 3D-points with the necessary accuracy. Movements parallel to the optical axis are ill conditioned for small FOV cameras, a stable triangulation of 3D points can hardly be achieved. An initial sideways camera movement is mandatory.

With additional metric depth information available it is possible to estimate metric 3D scene points and to establish a metric coordinate system for a single camera image. The coordinate system has its point of origin in the first camera center and the cameras optical axis is at zero rotation. Tracking can start from the first frame without an initial camera movement.

To facilitate tracking we estimate covariances for the 3D-points. The standard deviation of the PMD-depth data was measured by Kuhnert et al. in [8] as

$$\sigma_z = 2.734 * 10^{-3} d^2 + 2.867 * 10^{-3} d - 4.230 * 10^{-4},$$

d is the measured depth in meters. The standard deviation in x and y direction is influenced by the 2D feature tracking accuracy. In this work the KLT-corner-tracker [13] was used. Its standard deviation is given as $\sigma_{KLT} = 0.25$ pixel. This can be back-projected to the 3D-point by

$$\sigma_x = \sigma_y = \sigma_{KLT} / f_{PMD} * d$$

with d as above and f_{PMD} as the focal length of the PMD-camera in pixel.

Knowing the standard deviations along the three axes we are able to approximate the covariance matrix Σ_{3D}^{Cam} of a 3D-point X in the camera coordinate system. The transformation of Σ_{3D}^{Cam} into the global coordinate system Σ_{3D}^{World} needs the 4×4 affine transformation matrix T, with rotation matrix R and camera center C in global coordinates.

$$\Sigma_{3D}^{World} = T \Sigma_{3D}^{Cam} T^T \text{ with } \Sigma_{3D}^{Cam} = \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix} \text{ and } T = \begin{pmatrix} R & C \\ 0\ 0\ 0 & 1 \end{pmatrix}$$

3.2 Pose Estimation and Structure Update

By knowing 3D-points and a camera pose in a metric coordinate system it is now possible to estimate the camera movement for the next image in metric scale. 2D-features in the high resolution 2D-images can be tracked with high accuracy, and the 6DoF-camera pose can be estimated by a Levenberg-Marquardt-minimization from the known 2D/3D-correspondences, facilitated by a RANSAC to remove outliers [3].

Knowing the pose, again 3D-points and their covariances can be estimated in the global metric coordinate system as described above. Having two instances of the same 3D-point, these can be merged with a Kalman Filter [14]. Because the 3D-point is assumed to be static and its coordinates can be measured directly, the Kalman equations can be simplified to

$$\begin{aligned} X_{new} &= X_1 + G * (X_2 - X_1) \\ \Sigma_{X_{new}} &= (I_{3\times 3} - G) * \Sigma_{X_1} \end{aligned} \quad \text{with} \quad G = \Sigma_{X_2} * (\Sigma_{X_1} + \Sigma_{X_2})^{-1}$$

where $X_{1,2}$ are the 3D-positions of the points 1 and 2, $\Sigma_{X_{1,2}}$ are the 3×3-covariance matrices of points 1 and 2. X_{new} is the new merged point and $\Sigma_{X_{new}}$ its covariance. G is the gain matrix from the Kalman equations.

It is still possible to additionally use all SfM standard methods, i.e. triangulation of 3D-points in 2D-image parts without depth information. 3D-points can still be optimized by minimizing the back-projection error into the current 2D-image, a standard technique for SfM on 2D-images.

4 Results

The described method was evaluated on real as well as synthetic image sequences. The results are described in the following sections.

4.1 Synthetic Image Sequence

To be able to calculate correct pose estimation errors a synthetic 2D/3D-image sequence was used. For a description of the simulated camera setup see section 2.3, sample images are shown in figure 2. The image sequence is difficult to track for standard SfM, because it starts with a forward movement that is common i.e. for a moving robot but very disadvantageous for SfM.

The movement is starting in z-direction and the main movement is in the x-z-plane. It is overall covering $2.1m$ in x, $0.27m$ in y and $2.4m$ in z, the scene is at a distance of $4.5m$ to $7.5m$ for the starting image. The camera is also rotated during the movement. Figure 4 shows the camera path, the path is a closed loop of 101 images. Two sequences were generated, the first with ideal depth data despite the low spatial resolution, the second with noise added to the depth data. The σ_z of the noise was taken from [8].

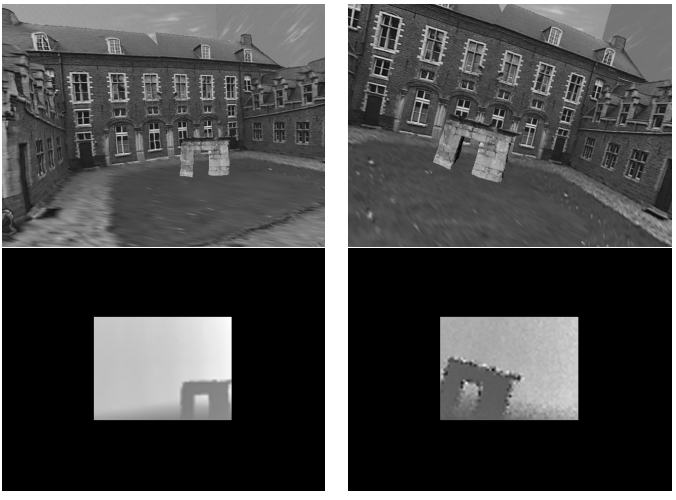

Fig. 2. Synthetic 2D/3D-image pairs. Left: no noise, Right: noise added.

SfM was run on these sequences evaluating three different scenarios:

1. 3D-points were created only from triangulation of 2D-features.
2. 3D-points were created only from the 3D-depth-images.
3. 3D-points were created from depth images as well as from 2D-triangulation, in regions where no 3D-information is available. This increases the solid angle in which known 3D-points are tracked to the FOV of the 2D-camera, while still keeping the metric initialization and the 3D-point stabilization.

For scenario 1 the whole sequence was scaled for comparison. The scale factor was calculated using the known ground truth distance for the initialization movement. In scenario 2 and 3 the estimated camera poses were not scaled or modified. The average translation and rotation errors are shown in table 1, absolute in m or degree as well as relative to the overall camera movement. The error progression over the sequences is shown in figure 3 for the ideal and noised depth data. Strong improvements are visible for scenarios 2 and 3 compared to scenario 1. Scenario 3 outperforms scenario 2 only for rotation estimation. The rotation in scenario 3 is stabilized by the higher FOV while translation estimation is worse due to the large z-errors of the triangulated points. On the sequence with the noise added, certain camera poses are estimated wrong. This does not introduce a drift, the pose estimation regenerates fast.

Please notice that in scenario 2 for the first 20 images all 3D-points are located in a solid angle of 40°, since all points are created from the narrow 3D-image. With the sideways camera movement 3D-points are then moving out of the PMD-camera FOV and points in a larger solid angle are tracked.

Figure 4 is showing the ground truth and the estimated camera paths as a top view on the x-z-plane. In figure 4(a) scenario 1 is shown. The main reason for the translational error is the drift that is accumulated over the whole sequence.

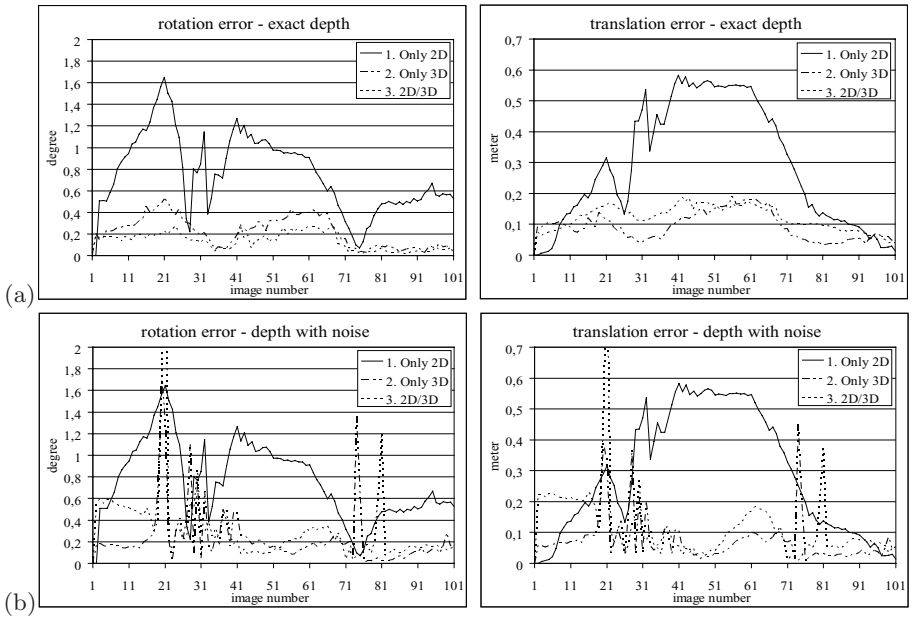

Fig. 3. Errors on synthetic sequence with (a) exact depth (b) noise on depth

Table 1. Average translation and rotation errors for synthetic sequences. The relative translation is relative to the total camera movement.

	Absolute Average Translation Error	Relative Average Translation Error	Absolute Average Rotation Error
1. Only 2D	$0.29m$	10.16%	$0.77°$
2. Only 3D	$0.09m$	2.63%	$0.22°$
3. 2D/3D	$0.12m$	3.41%	$0.13°$
2. Only 3D with noise	$0.08m$	3.37%	$0.23°$
3. 2D/3D with noise	$0.12m$	4.30%	$0.32°$

Its origin is the initial forward movement, that provides very bad triangulation baselines. The camera path for scenario 2 is shown in figure 4(b). Here the pose estimation for the forward movement is more accurate and thus the estimation overall better resembles the ground truth data.

4.2 Real Image Sequence

We used the 2D/3D-camera pair from section 2.3 generate an image sequence that is again starting with a forward movement of about 30 images and is very difficult for standard SfM. To be able to assess the quality on a free hand camera movement, 75 images were processed in a forward-backward-loop with a hard turn at image 75. Ideally the camera positions for forward and backward

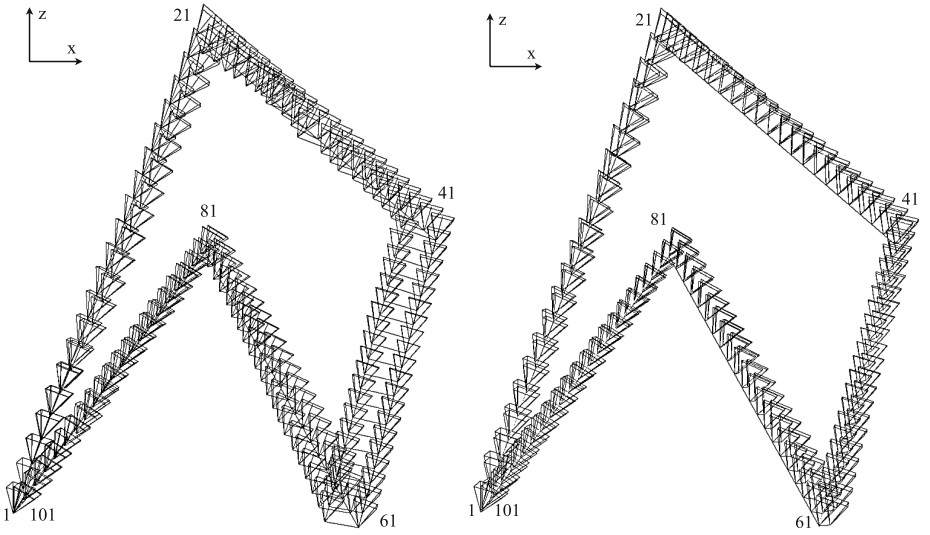

Fig. 4. Ground truth vs. estimated camera tracks on synthetic sequence without noise. (a) Scenario 1 - Only 2D (b) Scenario 2 - Only 3D.

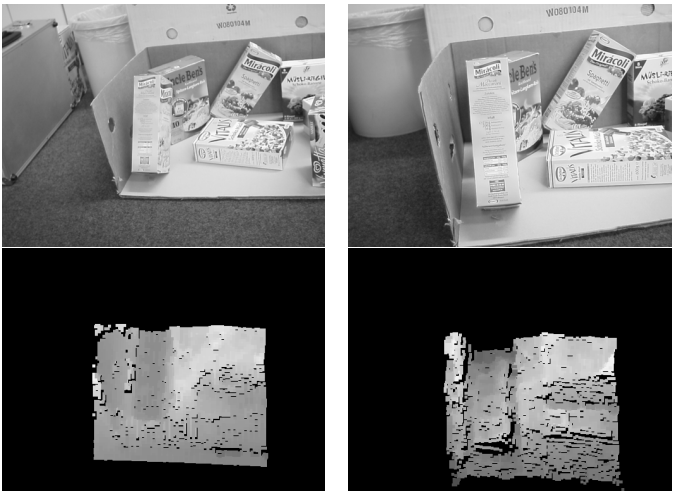

Fig. 5. 2D/3D-pairs from the real image sequence

movement should correspond. The scene distance is between $0.6m$ and $1.1m$, the camera movement spans $0.24m$ in x, $0.1m$ in y and $0.44m$ in z direction. Some images of the processed sequence and their depth maps are shown in figure 5.

The forward movement and the small 2D-camera FOV (42.3°) make SfM on 2D-images (scenario 1) impossible. While the movement direction is estimated

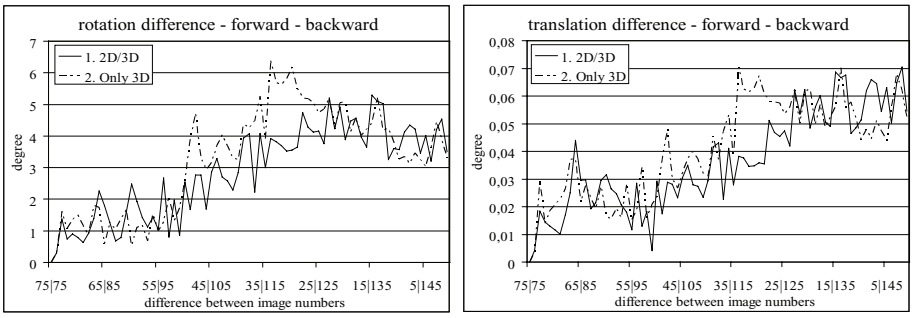

Fig. 6. Errors for the real image sequence. Differences between forward and backward track. Tracking on pure 2D data was impossible on this sequence.

correctly, the moved distance is estimated very inaccurate. The baseline for triangulating 3D-points is extremely short, thus the 3D-points have high errors in z-direction. This leads to a strong drift and a fast degradation of the camera track. No convergence could be reached.

The PMD-depth data is suitable to compensate for this distance misestimation. When using the 3D-data to aid pose tracking, SfM estimates a reasonable camera track. The average translation error between identical images in forward and backward movement for scenario 3 is $0.037m$, the average rotation error is $2.9°$. The overall performance on the real sequence is shown in figure 6. The error is small near image 75, where the forward-backward turn was just performed. Near image 1 the drift accumulated over 150 images under extremely difficult conditions. In contrast to the evaluation on synthetic data, scenario 3 slightly outperformed scenario 2 on real data, for translation as well as for rotation estimation. This was repeatable on different image sequences.

5 Conclusion

In this work it was shown how a 3D-PMD camera can be used to improve camera pose estimation with a modified SfM approach. The results show that the 3D-camera enables a valuable improvement to the camera tracking performance. Qualitatively because the estimation can be done starting from the first image in a metric coordinate system and quantitatively by improving the accuracy of the estimated pose.

Since PMD-cameras will experience a significant price drop in the next few years when productions processes are improved, such a camera provides an interesting and simple way to enhance the pose estimation of a single moving camera.

Acknowledgement. This work was supported by the German Research Foundation (DFG), KO-2044/3-1.

References

1. Bouguet, J.Y.: Camera calibration toolbox for matlab. `www.vision.caltech.edu/bouguetj/calib_doc/index.html` (1998)
2. Andrew J. Davison. Real-time simultaneous localisation and mapping with a single camera. In: Proc. ICCV (2003)
3. Hartley, R., Zissermann, A. (eds.): Multiple View Geometry in Computer Vision, 2nd edn. Cambridge university press, Cambridge (2004)
4. Koch, R., Koeser, K., Streckel, B., Evers-Senne, J.-F.: Markerless image-based 3d tracking for real-time augmented reality applications. In: WIAMIS 2005, Montreux, Switzerland (2005)
5. Koeser, K., Bartczak, B., Koch, R.: Drift-free pose estimation with hemispherical cameras. In: Proceedings of CVMP 2006, London (2006)
6. Kolb, A., Lindner, M.: Lateral and depth calibration of pmd-distance sensors. In: International Symposium on Visual Computing (ISVC06) (2006)
7. Kraft, H., Frey, J., Moeller, T., Albrecht, M., Grothof, M., Schink, B., Hess, H., Buxbaum, B.: 3d-camera of high 3d-frame rate, depth-resolution and background light elimination based on improved pmd (photonic mixer device)-technologies. In: 6 th Intl Conference for Optical Technologies, Optical Sensors and Measuring Techniques (OPTO 2004) (2004)
8. Kuhnert, K.-D., Stommel, M.: 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 (IROS06) (2006)
9. Lange, R., Seitz, P., Biber, A., Schwarte, R.: Time-of-flight range imaging with a custom solid-state imagesensor. In: EOS/SPIE Laser Metrology and Inspection, vol. 3823 (1999)
10. Lepetit, V., Fua, P.: Monocular model-based 3d tracking of rigid objects: A survey. Foundations and Trends in Computer Graphics and Vision 1(1), 1–89 (2005)
11. Pollefeys, M., van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J.: Visual modeling with a hand-held camera. International Journal of Computer Vision 59(3), 207–232 (2004)
12. Prasad, T.D.A., Hartmann, K., Wolfgang, W., Ghobadi, S.E., Sluiter, A.: First steps in enhancing 3d vision technique using 2d/3d sensors. In: Chum, V., Franc, O. (eds) Computer Vision Winter Workshop 2006, pp. 82–86, Telc, Czech Republic (2006)
13. Shi, J., Tomasi, C.: Good features to track. In: Conference on Computer Vision and Pattern Recognition, Seattle, June 1994, pp. 593–600. IEEE Computer Society Press, Los Alamitos (1994)
14. Welch, G., Bishop, G.: An introduction to the kalman filter. Technical Report TR 95-041, University of North Carolina, Department of Computer Science (2001)
15. Xu, Z., Schwarte, R., Heinol, H., Buxbaum, B., Ringbeck, T.: Smart pixel - photonic mixer device (pmd). In: M2VIP '98 - International Conference on Mechatronics and Machine Vision in Practice, pp. 259 – 264 (1998)