SLAM combining ToF and High-Resolution cameras

Victor Castañeda Diana Mateus Nassir Navab
Computer Aided Medical Procedures (CAMP)
Technische Universität München, Germany
http://campar.in.tum.de/

Abstract

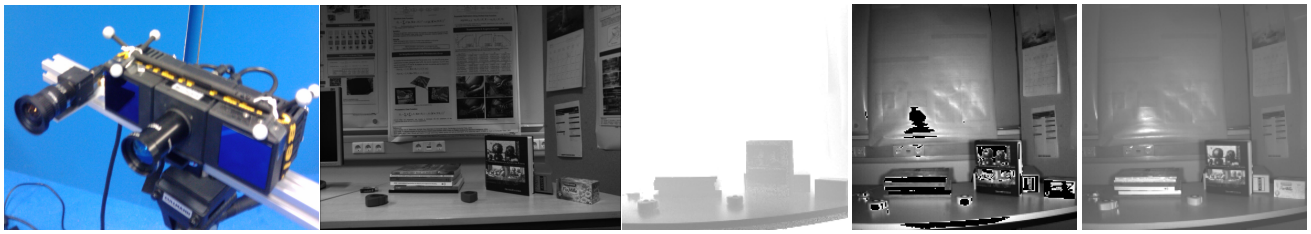
This paper describes an extension to the Monocular Simultaneous Localization and Mapping (MonoSLAM) method that relies on the images provided by a combined high resolution Time of Flight (HR-ToF) sensor. In its standard formulation MonoSLAM estimates the depth of each tracked feature as the camera moves. This depth estimation depends both on the quality of the feature tracking and the previous camera position estimates. Additionally, MonoSLAM requires a set of known features to initialize the scale of the map and the world coordinate system. We propose to use the combined high resolution ToF sensor to incorporate depth measures into the MonoSLAM framework while keeping the accuracy of the feature detection. In practice, we use a ToF (Time of Flight) and a high-resolution (HR) camera in a calibrated and synchronized set-up and modify the measurement model and observation updates of MonoSLAM. The proposed method does not require known features to initialize a map. Experiments show first, that the depth measurements in our method improve the results of camera localization when compared to the MonoSLAM approach using HR images alone; and second, that HR images are required for reliable tracking.

1. Introduction

The simultaneous localization and mapping (SLAM) problem consists of finding the position of an object (e.g. robot, camera, etc.) in a map, while simultaneously building the map as the object moves [9, 5]. Although SLAM has been widely studied in robotics and computer vision, it remains challenging due to the ill-posed nature of the problem, especially when online performance is desired. A large variety of SLAM approaches have been proposed that differ both in the type of sensors used (e.g onboard laser scanners [26, 22] or cameras [8, 19, 12, 21]) and in the actual algorithms used to estimate the map and sensor positions [23]. Given their availability and flexibility, the use of cameras has become very popular. One reference algorithm that solves SLAM from image data is the Monocular SLAM (MonoSLAM) proposed by Davison *et al*. [8]. In the MonoSLAM approach, the map is composed of 3D features that are estimated (up to scale) from a calibrated camera and image correspondences in two consecutive views. The estimation of the map and the camera position are alternated. The problem is formulated in terms of two dynamic systems that model the motion of the camera, the measurement (imaging) process and the noise. To find the current state of the systems an Extended Kalman filter is used. Several extensions to the original method have been proposed, for example, that employ stereo cameras [17, 20] to recover a more precise estimation of the 3D position of the features. In this paper, we investigate an extension to the MonoSLAM algorithm for a combined High-Resolution Time of Flight (HR-ToF) camera. As opposed to the above stereo approaches, the use of direct depth measures provided by the ToF camera enable our method to work under non-textured surfaces and poor lighting conditions.

In the last few years, there have been increasing advancements in the development of Time of Flight (ToF) cameras. ToF devices have faster frame rates ($\sim$40 fps) than laser scanners ($\sim$2 fps) [26, 22] and are therefore an interesting option to estimate 3D maps from a moving sensor. Recently, methods have been proposed that use ToF technology to estimate the pose of the camera in order to create 3D maps [6, 16, 24, 18]. For the most part these algorithms work off-line or use only the ToF camera information. We are instead interested in an online SLAM approach that takes advantage of the ToF high frame rates. Unfortunately, current ToF devices have low-resolution and precise feature tracking for online SLAM solutions such as that in [8] cannot be achieved.

To cope with the low resolution, new cameras that capture both color intensities and depth information per pixel [1, 2, 3, 4] are currently under development, though no yet commercially available. The combined sensor is often simulated using a ToF camera and a standard high-resolution RGB camera in stereo set-up [16, 11]. Calibration of such set-up suffices to provide RGB-d images. In



a) HR-ToF sensor b) High-resolution image c) ToF depth image d) ToF amplitude image e) ToF offset image

Figure 1: High-resolution ToF camera and the four images simultaneously captured per frame.

order to achieve real-time performance while maintaining high quality results, we propose to extend the MonoSLAM algorithm to use such an integrated HR-ToF sensor, composed of a ToF and a regular high-resolution (HR) camera. After the registration step we are able to incorporate the real-time depth information of the ToF camera in to the MonoSLAM framework, while reliably tracking the features in the high resolution camera. The result of the proposed HR-ToF SLAM are a 3D sparse map in metric coordinates (no longer up to-scale) and the trajectory of the camera.

The experiments in section 5 show the performance of the proposed method applied to the HR-ToF in comparison to the MonoSLAM approach applied to high-resolution images (HR-SLAM), and to our method using only the low-resolution offset images of the ToF camera. In particular, we measure the camera localization error w.r.t. the ground truth position of the camera obtained with an optical motion capture system. As expected, the availability of direct depth measures improves the localization precision and accuracy, while the high-resolution image guarantees a long-term tracking.

In the section 2 we recall the functioning principles of the ToF cameras and briefly describe the registration procedure to simulate the combined HR-ToF sensor. The problem statement and method are explained in sections 3 and 4 respectively.

2. Combined HR-ToF Sensor

A Time of Flight (ToF) camera emits an intensity modulated sinusoidal signal of infrared (IR) light and uses a CCD/CMOS sensor to detect the reflected light. According to the time-of-flight principle, the phase between emitted and received sinusoidal signals is proportional to the distance between the light source and the reflecting surface. Thus, by measuring the phase, amplitude and the offset of the received signal is possible to calculate a depth value for each pixel. Given that the signal is periodic, the modulation frequency limits the maximum distance which the camera can measure to half period of the modulated signal (e.g. from 50cm to 7.5m) [7, 13].

The ToF camera provides at least two images: the depth (Fig.1-c) and the amplitude (Fig.1-d). The *depth image* gives the measured distance between the reflecting surface and each pixel in the sensor. The *amplitude image* captures at every pixel the amount of IR light that was reflected back to the camera. The amplitude serves as a quality measure of the depth, as poor quality measures arise from low-amplitudes. In addition, the ToF camera may also provide an *offset image* (Fig.1-e), which is usually used as gray-scale image. The amplitude and the offset images can be used to calibrate the intrinsic and external parameters of the camera as well as the distortion parameters. The calibration procedure is equivalent to the used for high-resolution cameras, *e.g.* moving a known calibration pattern (chessboard) in front of the camera and optimizing the intrinsic and distortion parameters over a video sequence.

In order to simulate the behavior of a combined sensor capable of measuring reflected light and depth, a high resolution and a ToF cameras are mounted in a rigid stereo set-up, as illustrated in Figure 1-a. The relation between the two views is then found by performing a standard stereo calibration with the amplitude or offset image, moving a checkerboard in front of the cameras. To give a depth estimate to each pixel in the high resolution image, we use an inverse weighted distance interpolation. This method gives a depth value $\lambda(p)$ to pixel with coordinates p based on a weighted average of M available depth values λ_i at neighboring positions p_i $(1 \leq i \leq M)$:

$$\lambda(p) = \sum_{i=0}^{M} \frac{w_i(p)}{\sum_{j=0}^{M} w_j(p)} \lambda_i. \quad (1)$$

where weights w_i are computed as $w_i(p) = \frac{1}{\text{dist}(p,p_i)^d}$ (we use $d = 1$). The number of neighbors M is determined by the number of depth values being projected to a fix window size around p, 30×30 in our experiments. This interpolation allows us to handle the uneven distribution of the depth values when projected to the high resolution image.

3. Problem Statement: Feature-based SLAM

Our goal is to solve the SLAM problem using the combined HR-ToF sensor described above. We follow the standard formulation of feature-based SLAM based on dynamic

systems modeling the motion of the sensor and the measurement process. In [8], image measurements are obtained by establishing feature correspondences in the images over time. Depth measurements are computed using the camera geometry (calibration) and its estimated motion. We introduce an extension, where the depth estimates are directly obtained from the combined HR-ToF sensor. The extension includes the depth in the state vector, and modifies the measurement model and the observation update (see details in section 4). In the following we state the SLAM problem formally and explain the extension to depth measures.

SLAM is the problem of localizing an object (here a camera) in a map that is being simultaneously built. In the dynamic system formulation of SLAM [9, 5, 8], the position of both the sensor (camera) x and the map Y are modeled to be state vectors that evolve over time. The map is considered to be a collection of N features, whose 3D positions at time k are denoted $y_{i,k}$ and grouped in a vector:

$$Y_k = \begin{pmatrix} y_{1,k} & y_{2,k} & \dots & y_{N,k} \end{pmatrix}^\top. \tag{2}$$

The state vector of the camera x_k in the k-th frame is composed of the camera 3D position r^w, orientation (in quaternions) q^{wc}, velocity v^w and angular velocity ω^c, that is:

$$x_k = \begin{pmatrix} r^w & q^{wc} & v^w & \omega^c \end{pmatrix}^\top, \tag{3}$$

where superscripts indicate the world (w) and camera (c) coordinate systems.

The evolution of the two state vectors, the camera position x_k and the map Y_k, is modeled in terms of two dynamic systems. First, the *camera motion model* that determines the state vector of the camera x_k at instant k by means of a function f. Second, the *measurement model* that models the measurement process by means of a function h that relates the actual feature positions Y_k to their measurements $Z_k = \begin{pmatrix} z_{1,k} & z_{2,k} & \dots & z_{N,k} \end{pmatrix}^\top$. The two dynamic systems take the form:

$$x_k = f(x_{k-1}, u_k) + w_k, \tag{4}$$
$$Z_k = h(x_k, Y) + v_k. \tag{5}$$

Eq. 4 predicts the camera position at time k as a function of its previous state x_{k-1}, an optional input u_k and a motion disturbance w_k modeled with a zero-mean Gaussian distribution with covariance Q_k. In Eq. 5, $h(x_k, Y)$ models the measurement process, and v_k is the measurement disturbance modeled again with a zero-mean Gaussian distribution with covariance R_k. Notice that in the case of the combined HR-ToF sensor, h needs to take account of the depth measures, as explained in section 4.

Given the above formulation, the SLAM problem becomes that of finding the estimates to the full state vector, composed of the camera and the map state vectors $[\hat{x}_k \ \hat{Y}_k]^\top$ where $\hat{}$ is used to denote variable estimates.

4. Proposed Method: HR-ToF SLAM

One common solution to the problem is to estimate the state vector by means of an Extended Kalman Filter (EKF) [27]. Kalman based algorithms compute the estimate of the state vector describing the camera $\hat{x}_k$ and the feature positions $\hat{Y}_k$ in two recursive steps: a prediction (time-update) and a correction (measurement-update). As an auxiliary outcome of the Kalman filtering, a covariance matrix P is obtained representing the uncertainty of each estimation:

$$P = \begin{pmatrix} P_{xx} & P_{xY} \\ P_{Yx} & P_{YY} \end{pmatrix}. \tag{6}$$

In sections 4.1 and 4.2 we describe the updates, the instantiation of the motion-model and the extension introduced to the measurement model in order to consider the combined HR-ToF images.

4.1. Time Update

As described in [8], the *time-update* upgrades the camera position at time k given conditions at time $k-1$. More precisely, the updated state of the camera $\hat{x}_{k|k-1}$ is first predicted based on its motion during previous frames. The covariance matrix corresponding to the camera position P_{xx} is updated accordingly. The time update is resumed in the following equations:

$$\hat{x}_{k|k-1} = f(\hat{x}_{k-1|k-1}, u_k), \tag{7}$$
$$P_{xx,k|k-1} = \nabla f \cdot P_{xx,k-1|k-1} \cdot \nabla f^\top + Q_k, \tag{8}$$

where ∇f is the Jacobian of function f evaluated at time $k-1$. The explicit dynamic model of the camera motion is $\hat{x}_{k|k-1} = f(\hat{x}_{k-1}, u_k) =$

$$\begin{pmatrix} r^w_{k|k-1} \\ q^{wc}_{k|k-1} \\ v^w_{k|k-1} \\ \omega^c_{k|k-1} \end{pmatrix} = \begin{pmatrix} r^w_{k-1} + (v^w_{k-1} + \Delta v^w) \cdot \Delta t \\ q^{wc}_{k-1} \times q((\omega^c_{k-1} + \Delta\omega^c) \cdot \Delta t) \\ v^w_{k-1} + \Delta v^w \\ \omega^c_{k-1} + \Delta\omega^c \end{pmatrix}, \tag{9}$$

where $q((\omega^c_{k-1} + \Delta\omega^c) \cdot \Delta t)$ is the angle change due to angular velocity in the quaternion representation. The velocities change per time step Δt is modeled as:

$$\Delta v^w = a^w \cdot \Delta t, \tag{10}$$
$$\Delta\omega^c = \alpha^c \cdot \Delta t, \tag{11}$$

where the acceleration a^w and the angular acceleration α^c are modeled as processes of Gaussian distribution and zero mean (refer to [8] for details).

4.2. Observation Update

The observation-update describes the new position of both the camera $\hat{x}_{k|k}$ and the map $\hat{Y}_k$ given newly observed feature positions Z_k in the combined HR-ToF images at time k, and the updated $\hat{x}_{k|k-1}$. The positions

are upgraded according to the error in the prediction $Z_k - h(\hat{x}_{k|k-1}, \hat{Y}_{k-1})$ and the optimal Kalman gain matrix W_k:

$$\begin{bmatrix} \hat{x}_{k|k} \\ \hat{Y}_k \end{bmatrix} = \begin{bmatrix} \hat{x}_{k|k-1} \\ \hat{Y}_{k-1} \end{bmatrix} + W_k \left[Z_k - h(\hat{x}_{k|k-1}, \hat{Y}_{k-1}) \right], \quad (12)$$

$$P_{k|k} = P_{k|k-1} - W_k \cdot S_k \cdot W_k^\top, \quad (13)$$

where S_k is the innovation or residual covariance matrix representing the uncertainty of the prediction at time k (see section 4.2.1 for computation details). Notice that here, Z_k is the collection of measurements obtained with the combined HR-ToF sensor, *i.e.* $z_{i,k} = (u_i \ v_i \ \lambda_i)^\top$, where λ_i is the depth value at image coordinates u_i, v_i, and $1 \leq i \leq N$.

In order to relate the 3D features Y_k to the image (2D) and depth measurements contained in Z_k, the pinhole camera model with an invertible distortion model is used. The position of the features in the image plane is computed by projecting the expected positions $h(\hat{x}_{k|k-1}, \hat{Y}_{k-1})$ of the 3D features Y_{k-1} given the updated camera position $\hat{x}_{k|k-1}$ and using the intrinsic and distortion parameters. Let h_i^c be the result of the prediction $h(\hat{x}_{k|k-1}, \hat{y}_{i,k-1})$ in the camera coordinate system for a feature i, with $1 \leq i \leq N$, then the coordinates of the projection of $\hat{y}_{i,k-1}$ are:

$$u_u = u_0 - K_u \cdot \frac{h_{i_x}^c}{h_{i_z}^c}, \qquad v_u = v_0 - K_v \cdot \frac{h_{i_y}^c}{h_{i_z}^c}, \quad (14)$$

where (u_u, v_u) is the image coordinates of the features without distortion, (u_0, v_0) are the coordinates of the principal point, and (K_u, K_v) are the focal lengths in each direction. An invertible distortion model [8] is used, so that the distorted (real) image positions (u_d, v_d) are found with the expressions:

$$u_d = \frac{u_u - u_0}{\sqrt{1 + 2 \cdot k_1 r^2}} + u_0, \qquad v_d = \frac{v_u - v_0}{\sqrt{1 + 2 \cdot k_1 r^2}} + v_0, \quad (15)$$

where $r = \sqrt{(u_u - u_0)^2 + (v_u - v_0)^2}$ is the radial distance from the center of the image using undistorted coordinates.

Once the coordinates of the projections of each feature are computed, the error between the expected and the measured positions $Z_k - h(\hat{x}_{k|k-1}, \hat{Y}_{k-1})$ are used to correct the Kalman filter prediction applying Eq. 12 to update the camera position $\hat{x}_k$ and map $\hat{Y}_k$.

The difference to the state vectors and measurements of [8] is summarized in the following table:

MonoSLAM	HR-ToF SLAM
$z_{i,k} = (u_i \ v_i)^\top$	$z_{i,k} = (u_i \ v_i \ \lambda_i)^\top$
$h_i = (u_{d_i} \ v_{d_i})^\top$	$h_i = (u_{d_i} \ v_{d_i} \ \lambda_i)^\top$

where $\lambda_i = \sqrt{{h_{i_x}^c}^2 + {h_{i_y}^c}^2 + {h_{i_z}^c}^2}$ is the depth of the image position (u_i, v_i) associated to feature i.

4.2.1 Innovation Formula

The innovation formula is used to update (correct) the Kalman filter model, that is to calculate the Kalman Gain W_k and the innovation S_k in Eq. 12 and Eq. 13. The correction is computed from the difference between the measured and predicted positions, the Jacobians of the projections ∇h and the measurement noise R_k. Recall from Sect. 4.2 that the innovation covariance matrix S_k represents the uncertainty of the predictions ($x_{k|k-1}$ and the set of h_i) at time k and depends on the uncertainty of both the camera x_k and map Y_k (contained in P) as well as the measurement noise R_k. The updates to the EKF are computed as follows:

$$W_k = P_{k|k-1} \cdot \nabla h^\top \cdot S_k^{-1}, \quad (16)$$

$$S_k = \nabla h \cdot P_{k|k-1} \cdot \nabla h^\top + R_k. \quad (17)$$

Given the block form of P in Eq. 6, the 3×3 sub-matrices of S_k associated to feature i, *i.e.* $S_{k,i}$ can be obtained using:

$$S_{k,i} = \frac{\delta h_i}{\delta x} P_{xx} \frac{\delta h_i}{\delta x}^\top + \frac{\delta h_i}{\delta x} P_{xy_i} \frac{\delta h_i}{\delta y_i}^\top + \frac{\delta h_i}{\delta y_i} P_{y_i x} \frac{\delta h_i}{\delta x}^\top + \frac{\delta h_i}{\delta y_i} P_{y_i y_i} \frac{\delta h_i}{\delta y_i}^\top + R_{k,i}. \quad (18)$$

To calculate $R_{k,i}$ we model the noise associated to the measurement of each feature. The model takes into account that several sources of noise affect the HR and ToF cameras. First, we model the noise of the HR image position R_{uv} with a linear radial function to take account the radial distortion, such that the noise increases when the measurement is further away from the center of the image. The image position noise model is then:

$$R_{uv} = \sigma_{uv}^2 \cdot \left(1 + \frac{r}{\max r}\right), \quad (19)$$

where σ_{uv} determines the minimum level of noise (here assigned to the image center) and the factor $\frac{r}{\max r}$ controls the increase in the noise factor as the coordinates u and v go away from the center.

Second, the noise of the measurement depth measurements is modeled as a linear function with respect to the depth itself, as far away depth measures are usually noisier. Similar to 19 the model also takes into account the radial error. However, this time the linear model reflects the concentration of the infrared light in the middle of the image that causes noisier measurements in the borders of the image. Additionally, we use the amplitude image provided by the ToF camera as final indicator of the measurement noise, as it is known that measurements computed from low amplitudes (when the amount of reflected light is low) are noisier. In practice, we employ a function $\sigma_\lambda = \sigma_{min} + (1 - A_{uv})$, where σ_{min} ensures a minimum level of noise and A_{uv} is the corresponding amplitude value. Thus, the depth noise

model is resumed in the expression:

$$R_\lambda = \sigma_\lambda^2 \left(1 + \cdot \frac{\lambda}{\max \lambda} \cdot \frac{r}{\max r}\right). \quad (20)$$

Finally, we assume independence between the noise of the depth information and the image position. The resultant noise matrix R has the form:

$$R_{k,i} = \begin{pmatrix} R_{uv} & 0 & 0 \\ 0 & R_{uv} & 0 \\ 0 & 0 & R_\lambda \end{pmatrix} \quad (21)$$

Using the proposed noise model, the innovation matrix S_k can be computed. In particular, the 3×3 block elements of the Jacobian in Eq. 18 take the form:

$$\frac{\delta h_i}{\delta y_i} = \begin{pmatrix} \frac{\delta u_u}{\delta y_{i_x}} & \frac{\delta u_u}{\delta y_{i_y}} & \frac{\delta u_u}{\delta y_{i_z}} \\ \frac{\delta v_u}{\delta y_{i_x}} & \frac{\delta v_u}{\delta y_{i_y}} & \frac{\delta v_u}{\delta y_{i_z}} \\ \frac{\delta \lambda}{\delta y_{i_x}} & \frac{\delta \lambda}{\delta y_{i_y}} & \frac{\delta \lambda}{\delta y_{i_z}} \end{pmatrix}. \quad (22)$$

The first two rows in the expression above are the same used in [8], however the third row contains the derivatives of the new measurement value λ, namely:

$$\frac{\delta \lambda}{\delta y_{i_x}} = \frac{y_{i_x}}{||y_i||}, \qquad \frac{\delta \lambda}{\delta y_{i_y}} = \frac{y_{i_y}}{||y_i||}, \qquad \frac{\delta \lambda}{\delta y_{i_z}} = \frac{y_{i_z}}{||y_i||}$$

where, $||y_i|| = \sqrt{y_{i_x}^2 + y_{i_y}^2 + y_{i_z}^2}$. Using this noise model and computing W_k and S_k according to Eqs. 16 and 17, the EKF is updated and the system is ready for iteration $k+1$.

4.3. Feature extraction and depth initialization

In standard monocular approaches to SLAM, the initialization of the features depth is difficult, as it is not possible to measure the real distance from a feature to the camera. The usual way to initialize the features is to fix the depth of a reduced set of features, and to find a depth estimate for any other feature taking into account the motion of the camera. In the particular case of MonoSLAM [8], the depth of a new feature is initialized as a probabilistic uniform distribution on a finite ray passing through both the projection of the feature in the image plane and the origin of the camera coordinate system. This probability distribution is updated with each new frame using a particle filter that weights each discrete value of depth according to the intersection of the current and previous backprojected rays, under the estimated motion of the camera. In the case of PTAM [12], an initial translational motion is used to recover the depth of the initial features using stereo disparities.

In our case the initialization becomes easier as the probabilistic estimation is simply replaced with the depth measurement provided by the ToF device in the location where image features are detected. We use SIFT [15] to detect the features in the high-resolution image and use the first position of the camera as the origin of the world coordinate system.

5. Experimental Validation

We recorded a video sequence of the camera moving in an office environment, where the distance of the camera to the objects ranges from 1 to 3 meters. We used a Point-Grey Flea2 HR camera (640×480) and a PMD CamCube2 (204×204) ToF camera in a stereo setup as shown in Figure 1-a. The cameras are calibrated and synchronized.

In order to validate the performance of our approach we measure the error of the camera position. To capture the ground truth trajectory for the evaluation we place infrared markers on the camera and track them (using a commercial Tracking system). The markers are registered to the camera coordinate system with an additional calibration step that uses a checkerboard equipped with infrared markers. The resultant ground truth trajectory provides the position and orientation of the camera at each frame.

We measure the error of the estimated camera trajectories with respect to their ground truth. Three trajectories are compared. The first is obtained with the proposed HR-ToF SLAM method. The second one is the result of our method when using only the ToF images (the HR image is replaced by the low-resolution offset image), here named ToF SLAM. Finally, the third trajectory is the estimated pose of the camera obtained with the MonoSLAM algorithm on the high resolution images (HR-SLAM). As shown in Fig. 2, the proposed HR-ToF SLAM method (blue line) follows very closely the ground truth (error near to zero). ToF SLAM (green line) rapidly looses track due to the difficulty of reliably tracking features in the low-resolution images and to infrared specular reflections. Finally, HR-SLAM (red line) is able to keep the track but leads to a less accurate trajectory. Error peaks observed in the HR-ToF SLAM and HR-SLAM are explained by fast camera movements. The overall mean and standard deviation errors (using the magnitude of the error in the three axis) over the sequence with 380 frames are reported in Table 1. The camera trajectory using the proposed extension with the combined HR-ToF camera has the lowest average error ($\sim$2.2 cms). Despite the low resolution of the offset images, our method applied to the ToF images along ($\sim$2.4 cms) still has lower average error compared to the MonoSLAM with the high-resolution camera ($\sim$4.6 cms). The 3D error in the graph is calculated at each frame in the camera coordinate system, in order to observe the error in the depth (Z_{axis}), and in the horizontal (X_{axis}) and vertical (Y_{axis}) axes.

To start building the maps we initialize the depth of the detected features in the first frame using the values provided by the ToF data (*c.f.* section 4.3). This is done for the initialization of the map in the three compared methods in order to avoid the need of knowing an object dimensions for the HR-SLAM. We manually select four features in the first image to create a common world coordinate system. The chosen features are required by the HR SLAM, as it needs

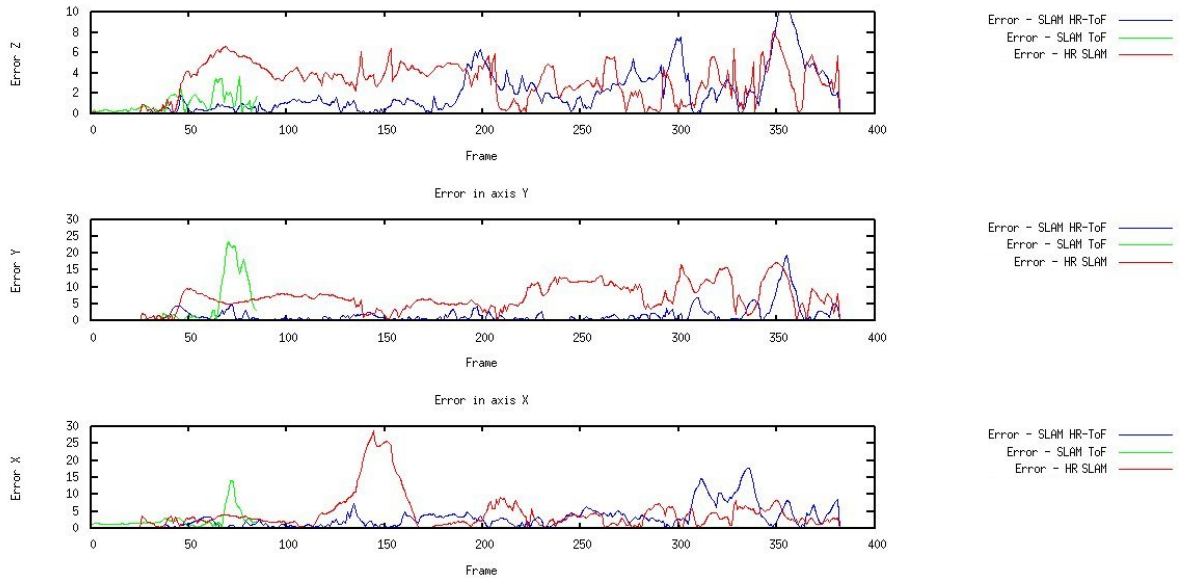


Figure 2: 3D error (X_{axis},Y_{axis} and Z_{axis}) in cms for each camera pose of the trajectory of the HR SLAM (red line), HR-ToF SLAM (blue line) and ToF SLAM (green line) versus the ground truth.

	HR-ToF (cms)	ToF (cms)	HR (cms)
X_{axis}	2.966 ± 3.312	2.426 ± 2.723	4.212 ± 5.187
Y_{axis}	1.632 ± 2.509	3.824 ± 6.628	6.657 ± 4.090
Z_{axis}	2.015 ± 2.190	1.018 ± 0.913	3.138 ± 1.895
Total	2.204 ± 2.670	2.450 ± 3.422	4.669 ± 3.724

Table 1: Mean and standard deviation of the camera localization error using the proposed approaches HR-ToF SLAM and ToF SLAM, and the monoSLAM algorithm applied individually to the high resolution (HR) camera.

a reasonable number of features for initialization. Although not required in our approach, we also provide the same four features to HR-ToF and to ToF SLAM for the sake of evaluation. Finally, we choose the world coordinate system to be the first frame position of the camera, this allows us to relate the camera coordinate system to the ground truth.

Snapshots in Fig. 3 show the difference between the uncertainty of the feature position estimates. The proposed HR-ToF method has less uncertainty than MonoSLAM applied to the high-resolution video. The difference is explained by the use of the ToF depth measurements.

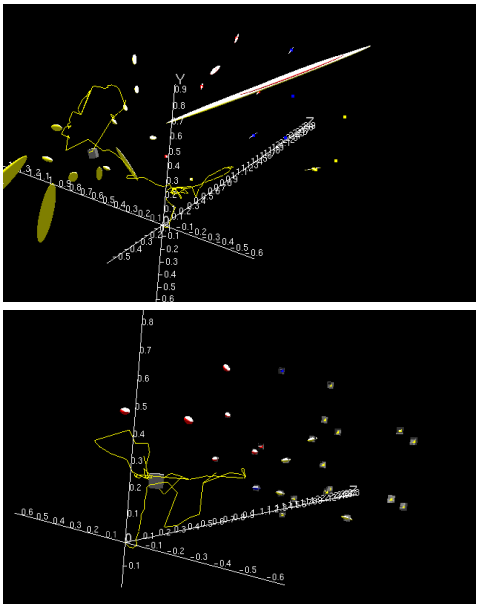


Figure 3: Snapshot of SLAM results for frame 380 of the sequence, based only on the HR camera (top) and on the combined HR-ToF sensor (bottom). The uncertainty in the depth data is displayed as ellipsoids. Lower uncertainties are achieved with the method HR-ToF SLAM.

6. Discussion and Future Work

We have presented a new online SLAM approach which uses a combined sensor gathering images from a high-resolution camera registered to a ToF device. The combined HR-ToF sensor includes valuable depth information while allowing for high-precision tracking. We address the SLAM problem by extending the measurement model and the innovation formulas of the MonoSLAM algorithm. Our

results show that these extensions improve the map uncertainty and the localization error of the camera.

Future work directions include detecting 3D features and using 3D tracking (scene-flow) methods to take full advantage of the depth image. In the short term, we would like to improve the registration using a combined stereo and depth calibration [10] method that models the depth error with B-spline or polynomial functions [14]. This is expected to reduce the noise of the depth measurements. Finally, it would be interesting to use the proposed model with an Unscented Kalman filter (UKF) [25] instead of Extended Kalman filter, since the UKF converges better when the measurements have a non-Gaussian distributed noise as is the case of the ToF depth measurments.

Acknowledgements This work was partially supported by the German Academic Exchange Service (DAAD) and Chilean National Commission for Science and Technology (CONICYT).

References

[1] http://pro.jvc.com/prof/attributes/features.jsp?model_id=MDL101309. 1

[2] http://www.primesense.com/. 1

[3] http://www.xbox.com/en-US/hardware/k/kinectforxbox360/. 1

[4] http://panasonic-electric-works.net/D-IMager/index.html. 1

[5] T. Bailey and H. Durrant-Whyte. Simultaneous localization and mapping (SLAM): part II State of the Art. *Robotics & Automation Magazine, IEEE*, 13(3):108–117, August 2006. 1, 3

[6] C. Beder, I. Schiller, and R. Koch. Real-time estimation of the camera path from a sequence of intrinsically calibrated pmd depth images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVII:45–50, 2008. 1

[7] B. Büttgen, T. Oggier, and M. Lehmann. CCD/CMOS lock-in pixel for range imaging: challenges, limitations and state-of-the-art. In *1st range imaging research day*, pages 21–32, 2005. 2

[8] A. Davison. Real-Time Simultaneous Localisation and Mapping with a Single Camera. In *IEEE International Conference on Computer Vision*, pages 1403–1410, October 2003. 1, 3, 4, 5

[9] H. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping: part I The Essential Algorithms. *Robotics & Automation Magazine, IEEE*, 13(2):99–110, June 2006. 1, 3

[10] U. Hahne and M. Alexa. Combining time-of-flight depth and stereo images without accurate extrinsic calibration. *Int. J. Intell. Syst. Technol. Appl.*, 5(3/4):325–333, 2008. 7

[11] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In *RGB-D: Advanced Reasoning with Depth Cameras Workshop in conjunction with RSS, Zaragoza, Spain*, 2010. 1

[12] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *IEEE International Symposium on Mixed and Augmented Reality*, Nara, Japan, 2007. 1, 5

[13] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight sensors in computer graphics. In *Proc. Eurographics (State-of-the-Art Report)*, 2009. 2

[14] M. Lindner and A. Kolb. Calibration of the intensity-related distance error of the PMD ToF-Camera. In *SPIE: Intelligent Robots and Computer Vision XXV*, volume 6764, pages 6764–35, 2007. 7

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision (IJCV)*, 60:91–110, 2004. 5

[16] S. May, D. Droeschel, D. Holz, S. Fuchs, E. Malis, A. Nüchter, and J. Hertzberg. Three-dimensional mapping with time-of-flight cameras. *Journal of Field Robotics, Special Issue on Three-Dimensional Mapping, Part 2*, 26(11-12):934–965, December 2009. 1

[17] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid. A constant time efficient stereo SLAM system. In *British Machine Vision Conference (BMVC)*, 2009. 1

[18] E. Menegatti, A. Zanella, S. Zilli, F. Zorzi, and E. Pagello. Range-only SLAM with a mobile robot and a wireless sensor networks. In *Int. Conf. on Robotics and Automation (ICRA)*, pages 1699–1705, 2009. 1

[19] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Conference on Artificial Intelligence (AAAI)*, pages 593–598, 2002. 1

[20] P. Mountney, D. Stoyanov, A. Davison, and G. Yang. Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. In *Int. Conf. on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, pages 347–354, 2006. 1

[21] R. A. Newcombe and J. Andrew. Live dense reconstruction with a single moving camera davison. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2010. 1

[22] A. Nüchter, K. Lingemann, J. Hertzberg, and H. Surmann. 6D SLAM - 3D mapping outdoor environments: Research articles. *J. Field Robot.*, 24(8-9):699–722, 2007. 1

[23] H. Strasdat, J. M. M. Montiel, and A. J. Davison. Real-time monocular SLAM: Why filter? In *Int. Conf. on Robotics and Automation (ICRA)*, pages 2657–2664, 2010. 1

[24] B. Streckel, B. Bartczak, R. Koch, and A. Kolb. Supporting structure from motion with a 3D-range-camera. In *Scandinavian Conf. Image Analysis (SCIA)*, LNCS, pages 233–242. Springer, 2007. 1

[25] N. Sunderhauf, S. Lange, and P. Protzel. Using the unscented kalman filter in mono-slam with inverse depth parametrization for autonomous airship control. In *Safety, Security and Rescue Robotics, 2007. SSRR 2007. IEEE International Workshop on*, pages 1 –6, 2007. 7

[26] R. Triebel and W. Burgard. Improving simultaneous localization and mapping in 3D using global constraints. In *Conference on Artificial Intelligence (AAAI)*, 2005. 1

[27] G. Welch and G. Bishop. An introduction to the Kalman filter. Technical report, Chapel Hill, NC, USA, 1995. 3