2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 13–16 October 2015, Banff, Alberta, Canada

Pose Estimation Using Monocular Vision and Inertial Sensors Aided with Ultra Wide Band

Hanna E. Nyqvist*, Martin A. Skoglund*, Gustaf Hendeby*†, and Fredrik Gustafsson*
* Dept. of Electrical Engineering, Linköping University, Linköping, Sweden
† Dept. of Sensor & E/W Systems, Swedish Defence Research Agency (FOI), Linköping, Sweden

***Abstract*—This paper presents a method for global pose estimation using inertial sensors, monocular vision, and *ultra wide band* (UWB) sensors. It is demonstrated that the complementary characteristics of these sensors can be exploited to provide improved global pose estimates, without requiring the introduction of any visible infrastructure, such as fiducial markers. Instead, natural landmarks are jointly estimated with the pose of the platform using a simultaneous localization and mapping framework, supported by a small number of easy-to-hide UWB beacons with known positions. The method is evaluated with data from a controlled indoor experiment with high precision ground truth. The results show the benefit of the suggested sensor combination and suggest directions for further work.**

I. Introduction

To provide full six *degrees of freedom* (DoF) pose estimates in indoor environments — *indoor positioning and navigation* (IPN) — in real time is an enabling ability for many applications. A typical example is augmented reality, where an accurate camera pose is needed to superimpose graphics on the image; *e.g.*, in TV productions [1, 2] and head-mounted displays [3]. The technique is also used for vehicle and cargo localization in warehouses [4]; localization of smartphone users [5]; and mobile robotics [6]. The systems available today predominantly rely on pre-installed infrastructure, *e.g.*, reference markers of some type, or detailed maps in combination with visual, infrared, ultra sound, or laser sensors, see *e.g.*, [7] for an overview. Adding infrastructure or providing maps of environments prone to change can be both time consuming and difficult to do. This limits the applicability of these systems.

Solutions that avoid the need for external infrastructure or pre-mapping of the environment using either *inertial measurement units* (IMUs), see *e.g.* [8], or cameras exist. Pure IMU solutions, *e.g.*, [9–11], where the inertial measurements in one way or another are integrated over time, suffer from drift already on relatively short term. Camera based systems can alleviate the need for pre-mapping and external infrastructure by online mapping of naturally occurring features in a *simultaneous localization and mapping* (SLAM) solution [12–14]. These solutions drift less than the IMU solutions, but have issues with long term drift and cannot determine the scale of the map, which is required to relate the map to physical dimensions.

A popular solution is to combine camera and IMU and use naturally occurring features to provide pose estimates [15–18]. The resulting SLAM solutions are however still unable to provide global pose estimates, and do furthermore tend to drift over time as a result of accumulating measurement errors over time [19]. Both problems can be handled by providing extra information in the form of a map [20]. However, to create the map can be time consuming and difficult. This paper therefore proposes to instead combine the visual-inertial SLAM solution with measurements of the distance to a few known anchor points using *ultra wide band* (UWB) radio signals. Including the UWB measurements makes is possible to provide globally consistent poses and alleviate the drift problem. Also, compared to providing a complete map, measuring the positions of a few UWB tags is easy. The usage of IMUs, cameras, and UWB systems for IPN is nothing new [21, 22], but the idea to fuse all the three of them is not well explored.

UWB measurements provide the distance to anchor points and can provide global drift-free positioning (however, only three DoF). The position accuracy has been reported to be in the range sub-millimeter to a few decimeters, see [23] for an overview. It has been suggested to fuse IMU and UWB measurements to improve pose estimation performance [24–26]. The papers [24, 25] use the sensor combination only to track position and ignores the problem of estimating the orientation. The paper [26] attempts full six DoF tracking, but experience problem in obtaining height estimates with accuracy comparable with the estimates in the plane. In [27] camera, IMU, and UWB sensors are all used together to track the position and velocity in the plane of a flying drone. Contrary to the method suggested in this paper, no attempt is made to obtain a full six DoF pose of the drone. So far, estimation of the full global six DoF pose using visual-intertial SLAM aided by UWB measurements of anchor points has not been attempted.

The problem considered in this paper is sequential estimation of the position and orientation of a platform moving in six DoF. The three platform mounted sensors considered are: an IMU, measuring its acceleration and rotational velocity relative to an inertial reference frame; a UWB radio signal receiver, measuring the distance to several stationary UWB beacons; and a monocular camera, producing angular measurements. A filter that fuses the information from all these three sensors is proposed and evaluated using real data. The results are then compared with filters using measurements from only one or two of the sensors.

The outline of this paper is as follows. The models and the filter designed to solve the IPN problem is given in Sec. II. The used sensors are then discussed in Sec. III. Experimental design and results are presented in Sec. IV, followed by concluding remarks in Sec. V.

978-1-4673-8402-5/15/$31.00

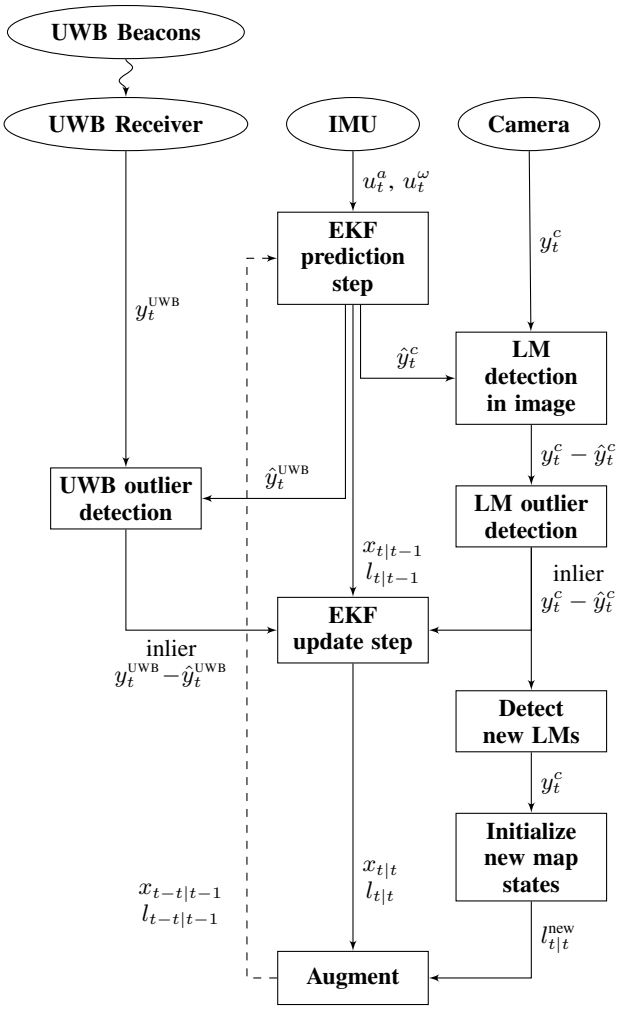


Fig. 1: Flowchart of the proposed tracking filter. The complete notation is introduced throughout Sec. II. Note that also uncertainty information in the form of covariance matrices are passed between the different steps in the flowchart but this is not explicitly stated in the figure.

II. System Models and Filtering Solution

An overview of the proposed pose estimation filter is provided in Fig. 1. The pose of the sensor platform and the visual landmarks are unknown but coupled through the sensor data. This results in a SLAM problem which has the typical form

$$x_{t+1} = f(x_t, u_t, w_t), \tag{1a}$$
$$l_{t+1} = l_t, \tag{1b}$$
$$y_t = h(x_t, l_t, e_t), \tag{1c}$$

where $f(\cdot)$ describes the platform dynamics, w_t is process noise, l_t are static visual landmarks, $h(\cdot)$ is the measurement function relating the states and the landmarks to the measurements, and e_t is measurement noise.

The following parts of this section describe how the sensor measurements and the motion of the system can be modeled.

A. Coordinate Systems

To be able to precisely describe the individual systems and the sensors, the following coordinate systems are defined:

- **Room fixed coordinate system, e** — This system has its origin, O_e, in the surroundings in which the sensor platform is moving and its axes are fix relative to the stationary environment. This frame is considered inertial.
- **IMU fixed/system fixed coordinate system, b** — This system is fixed relative to the sensor platform, which is to be tracked, and is referred to as the body frame. Its origin, O_b, is in the center of the accelerometer triad and is oriented such that it is aligned with the accelerometer and the gyroscope axes.
- **Camera fixed coordinate system, c** — This system has its origin in the optical center of the camera integrated on the IMU unit with the z-axis is pointing outwards from the camera parallel to the z-axis of the IMU.

For convenience the IMU and the platform frame b is assumed to coincide, avoiding all complicating lever-arm effects when using this sensor. The aim with the designed filter is hence to determine the position and the orientation of the platform/IMU b relative to the room e as a function of time.

B. Inertial Measurements and Dynamic Motion Model

An ideal strap-down accelerometer measures the platform referenced acceleration and gravitational field

$$y_t^a = a_t^b + g_t^b, \tag{2}$$

while an ideal gyroscope measures the rotational velocity

$$y_t^\omega = \omega_t^b. \tag{3}$$

In order to keep state dimension small the inertial measurements can be viewed as known signals according to

$$u_t^a = a_t^b + g_t^b, \tag{4a}$$
$$u_t^\omega = \omega_t^b. \tag{4b}$$

By neglecting sensor biases and drift the 10-state motion model in [28] can be used

$$x_{t+1} = \begin{bmatrix} p_{t+1}^e \\ v_{t+1}^e \\ q_{t+1}^{be} \end{bmatrix} = \begin{bmatrix} I_3 & TI_3 & 0 \\ 0 & I_3 & 0 \\ 0 & 0 & I_4 \end{bmatrix} \begin{bmatrix} p_t^e \\ v_t^e \\ q_t^{be} \end{bmatrix} + \begin{bmatrix} \frac{T^2}{2} I_3 & 0 \\ TI_3 & 0 \\ 0 & \frac{T}{2} \end{bmatrix} \begin{bmatrix} \mathcal{R}(q_t^{be})^T u_t^a - g^e \\ S(u_t^\omega) q_t^{be} \end{bmatrix} + \begin{bmatrix} \frac{T^2}{2} I_3 & 0 \\ TI_3 & 0 \\ 0 & \frac{T}{2} \widetilde{S}(q_t^{be}) \end{bmatrix} \begin{bmatrix} w_{a,t}^b \\ w_{\omega,t}^e \end{bmatrix}, \tag{5}$$

where

$$S(u_t^\omega) = \begin{bmatrix} 0 & -\omega_{x,t} & -\omega_{y,t} & -\omega_{z,t} \\ \omega_{x,t} & 0 & \omega_{z,t} & -\omega_{y,t} \\ \omega_{y,t} & -\omega_{z,t} & 0 & \omega_{x,t} \\ \omega_{z,t} & \omega_{y,t} & -\omega_{x,t} & 0 \end{bmatrix}, \tag{6a}$$

$$\widetilde{S}(q_t^{be}) = \begin{bmatrix} -q_{1,t} & -q_{2,t} & -q_{3,t} \\ q_{0,t} & -q_{3,t} & q_{2,t} \\ q_{3,t} & q_{0,t} & -q_{1,t} \\ -q_{2,t} & q_{1,t} & q_{0,t} \end{bmatrix}. \tag{6b}$$

Here $p_t^e = [x_t^e, y_t^e, z_t^e]^T$ is the position, $v_t^e = [v_{x,t}^e, v_{y,t}^e, v_{z,t}^e]^T$ the velocity, and $q_t^{be} = [q_{0,t}, q_{1,t}, q_{2,t}, q_{3,t}]^T$ a unit quaternion parametrizing the orientation of the platform. A rotation matrix $\mathcal{R}(q_t^{be}) = \mathcal{R}_t^{be} \in \mathbb{SO}(3)$ can be parametrized using the unit quaternion. $\mathcal{R}(q_t^{be})^T u_t^a - g^e$ is the specific force input expressed in the e frame, where $g^e \approx [0,\ 0,\ 9.81]^T$ is the local earth gravitational field. Note also that the noise model depends on the orientation of the platform.

C. Camera and Landmark Parametrization

The camera calibration is found using [29] which allows us to work with normalized image coordinates m_t^c. Landmarks are in the filter parametrized using Cartesian coordinates in the e-frame as $l_t^e = [l_{x,t}^e,\ l_{y,t}^e,\ l_{z,t}^e]^T$ while they are expressed in the c-frame using

$$l_t^c = \mathcal{R}^{cb}\mathcal{R}_t^{be}(l_t^e - p_t^e) + p_c^c, \tag{7}$$

where the static relative pose of the camera and the IMU, represented by $\mathcal{R}^{cb}$ and p_c^c, is estimated as in [30]. With the standard pin-hole projection, a camera measurement model is

$$y_t^c = \frac{1}{l_{z,t}^c}\begin{bmatrix} l_{x,t}^c \\ l_{y,t}^c \end{bmatrix} + e_t^c = h_c(x_t, l_t) + e_t^c, \tag{8}$$

which is a function of both the landmarks and the pose of the platform. We refer to this parametrization as the *direct parametrization* (DP).

A slightly different approach is proposed in [31] where the authors introduce the *inverse depth parametrization* (IDP), which simply decouples the camera orientation from the landmark position. IDP has a small linearization error even for large uncertainty in depth and it is easy to represent the range of depth uncertainty including infinity in a confidence region. This makes the parametrization suitable for points at unknown distance.

To achieve this, six parameters are used. The three first are the coordinates of the camera from which the landmark was first observed p_t^e. The remaining three parameters describe the vector from the camera to the landmark encoded by two angles, φ^e and θ^e, and the inverse depth, ρ^e,

$$l_t^e = p_t^e + \frac{1}{\rho_t^e} d(\varphi_t^e, \theta_t^e), \tag{9a}$$

$$d(\varphi_t^e, \theta_t^e) = \begin{bmatrix} \cos\varphi_t^e \sin\theta_t^e \\ \sin\varphi_t^e \sin\theta_t^e \\ \cos\theta_t^e \end{bmatrix}. \tag{9b}$$

The angles are computed from the normalized image coordinates as

$$g_t^e = \mathcal{R}_t^{eb}\mathcal{R}^{bc} m_t^c, \tag{10a}$$

$$\varphi_t^e = \arctan2(g_{y,t}^e, g_{x,t}^e), \tag{10b}$$

$$\theta_t^e = \arctan2(\|[g_{x,t}^e\ g_{y,t}^e]^T\|_2, g_{z,t}^e), \tag{10c}$$

where $\arctan2$ is the four-quadrant arctangent function and the inverse depth can be initiated with any positive number.

The corresponding measurement equation at time t for a landmark initiated at time j in the camera frame is then

$$l_t^c = \mathcal{R}^{cb}\mathcal{R}_t^{be}\big(\rho_t^e(p_t^e - p_j^e - \mathcal{R}_t^{eb}\mathcal{R}^{bc}p_c^c) + d(\varphi_t^e, \theta_t^e)\big), \tag{11a}$$

$$y_t^c = \frac{1}{l_{z,t}^c}\begin{bmatrix} l_{x,t}^c \\ l_{y,t}^c \end{bmatrix} + e_t^c. \tag{11b}$$

This parametrization is used in [16] for SLAM using monocular vision and inertial sensors. The same sensor setup was used in [15] showing that feature initialization and prediction in difficult cases, such as forward motion, can be handled better using IDP than DP with support of IMU.

D. UWB Measurements

The UWB measurements are in this paper modeled as the distance between the beacons b^i and the receiver r

$$y_t^{\text{UWB}^i} = \|p_{b^i}^e - p_{r,t}^e\|_2 + e_t^{\text{UWB}^i}, \quad i \in \{1, 2, \ldots, n^b\} \tag{12}$$

where

$$p_{r,t}^e = p_t^e + \mathcal{R}_t^{eb} p_r^b, \tag{13}$$

p_r^b is the position of the UWB receiver relative to the IMU and it is assumed to be known. Also $p_{b^i}^e$, the positions of the UWB beacons in the e-frame, are assumed to be known.

E. Filtering Solution

We can now put together all the models defined above, which then becomes a nonlinear filtering problem. Let the state vector, x_t consist of the motion model (5), and the set of landmarks, $\{l_k^e\}_{k=1}^{N_l}$ where N_l may vary over time. Together with the sensor models the SLAM system becomes

$$x_{t+1} = f(x_t, u_t^a, u_t^\omega) + B(x_t)w_t, \tag{14a}$$

$$l_{t+1} = l_t, \tag{14b}$$

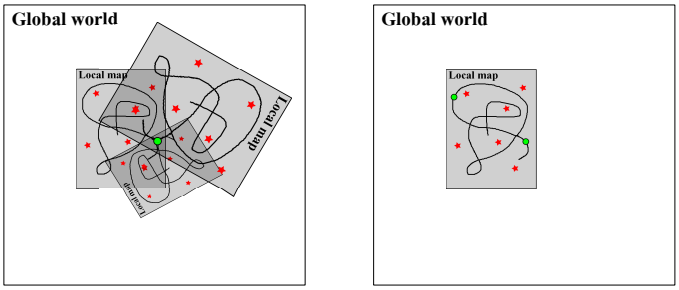
$$y_t^c = h_c(x_t, l_t) + e_t^c \tag{14c}$$

$$y_t^{\text{UWB}} = h_{\text{UWB}}(x_t) + e_t^{\text{UWB}} \tag{14d}$$

where the landmark dynamics are zero since these are assumed stationary and the coordinate frame superscript has been dropped. A simplistic, yet powerful, approach to the nonlinear filtering problem (14) is to apply EKF-SLAM [32] which is an EKF applied to a SLAM system model. The model (14) is first used in a prediction step where states and landmarks, $x_{t|t-1}$ and $l_{t|t-1}$ together with measurements, $\hat{y}_t^{\text{UWB}}$ and $\hat{y}_t^c$, are predicted. This prediction is based on measurements gathered only up to the previous time step. In a second step the predictions are adjusted, $x_{t|t}$ and $l_{t|t}$, based on measurements obtained at the current time step.

The measurement Jacobian is often sparse since the camera will typically only generate measurements for a subset of all landmark states at each time instant and an efficient implementation exploits this structure. Since the measurements are assumed independent, the measurement update can be processed iteratively avoiding the need for inverting a large matrix in the Kalman gain computation.

Note that a similar approach can also be used if some beacon locations, say $p_{b^j}^e\ j \in \{1, 2, \ldots, \}$, are unknown. This is then done by appending the unknown locations to the state vector with zero dynamics, $p_{b^j,t+1}^e = p_{b^j,t}^e$. Note that the UWB



(a) The local map with information about one global position can be rotated arbitrarily around this global position and has an unresolved scale.

(b) Having information about more global positions resolves the scale and rotational degree of freedom of the local map and fixes it in the global world.

Fig. 2: Illustration showing how the positioning setup with a camera and an IMU in combination with UWB TOA measurements gives global observability. The black line is a track, red stars are detected landmarks and green dots are global position information achieved from the UWB system.

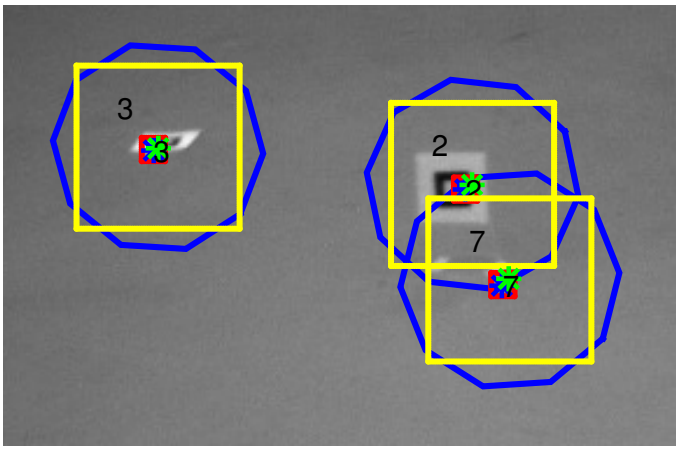


Fig. 3: An example from the image processing pipeline. The green stars represent the predicted locations of the landmarks and the blue ellipses correspond to the prediction uncertainty. The yellow boxes show the regions in which the NCC between the landmark descriptors and the image is computed. The red box show the patch with the NCC maximum. The blue stars represent the measurements forwarded to the filter.

measurements now depend on both the receiver and beacon positions

$$y_t^{\text{UWB}} = h_{\text{UWB}}(x_t, p_{b^j}) + e_t^{\text{UWB}}. \tag{15}$$

This of course makes the estimation problem much harder due to the increased degrees of freedom.

F. Resolving a Global Estimate

Combining vision, IMU, and UWB in a SLAM system has some appealing properties. A monocular camera can alone only provide locally consistent motion and map estimates defined up to an unknown scale. That is, it is impossible to distinguish the scale of the scene from motion, or in other words, if the scene is far away and the motion is fast or if the scene is close and the motion is slow. An IMU can help resolve this issue with scale but globally consistent estimates are still not achievable without for example information about the exact initial pose of the platform. The UWB does both resolve the unknown scale and admits estimation of the absolute pose and map in the e frame. The global position of the locally consistent map and motion obtained with a the camera or camera-IMU setup can be inferred from the UWB distance measurements. Also the rotation relative to the global frame can be determined without any knowledge of the initial rotation since two distinct globally known points (not aligned with gravity), *e.g.* two beacons, and the direction of gravity uniquely determines the orientation. This sub-map to global reference is illustrated in Fig. 2.

III. Sensor Properties

In this section we discuss some properties of the sensors used in this paper and how obtained measurements are processed in order to fit into the filtering framework described in Sec. II-E.

A. Image Processing

The image processing pipeline is coupled with state estimation allowing tight search regions and outlier rejection using the filter residuals. Features are generated based on the Harris corner detector [33] which can be computed in image regions to limit the computational complexity.

The 10 strongest Harris responses which are above a threshold, are selected with the condition that they need to be separated with at least 40 pixels. Around each detection a 11×11 pixel patch is extracted representing a local feature descriptor. Since an estimate of the current pose and all landmarks are available in the filter the measurements can be predicted using (8). Then, the *normalized cross-correlation* (NCC) between the patch and a 25×25 pixel search region, centered at the prediction, is computed for each landmark that is predicted to be within the field of view. The maximum from the NCC operation is considered to be the measurement in the EKF-SLAM filter. See Fig. 3 for a crop of a full image from the scene illustrating, predictions, matches, uncertainty ellipsoids and search-regions. This is a rather simple approach which does not account for occlusion and does not handle perspective changes. The highest response from NCC is taken as the measurement. Outliers are rejected using the norm of the measurement residuals normalised by the inverse square root of the innovation covariance

$$r^i = \|S^{-1/2}(y^i - \hat{y}^i)\|^2. \tag{16}$$

These r^i's are χ^2-distributed based on the assumption that the residuals are zero-mean Gaussian and are rejected if $r_i > 9$ corresponding to a 99% confidence. This gating approach should not be too conservative since the depth of newly initialized landmarks may be far off and thus result in large residuals.

The algorithm searches for new features as soon as the number of landmarks which are predicted to be within the

field of view drops below 10.

Loop closures do not work well based on the prediction of landmark locations due to two reasons; for longer loops the integrated drift will be too large resulting in unreliable loop closure candidates; the pixel patch descriptor is not invariant to perspective changes. In future solutions more sophisticated features, such as SIFT [34], should be used as it also enables appearance based loop closures [35].

B. UWB Processing

UWB is a radio technology using high bandwidth signals and has applications in fields like communications, localization, and radar. The technology typically makes use of very short transmitted pulses to achieve the high bandwidths. Using high bandwidth signals has the advantage of giving high spatial resolution which makes accurate positioning possible. In this paper we make use of *time of arrival* (TOA) measurements obtained from a UWB system. A good overview of the UWB TOA based challenges and techniques is given in [36]. Here we will give a short summary.

TOA measurements for a signal transmitted from a beacon b to a receiver r can under ideal *line-of-sight* (LOS) conditions be modeled as

$$\tau = \frac{\|p_r - p_b\|_2}{c} + \Delta\tau + e_\tau \tag{17}$$

as done in [37]. Here p_r is the position of the receiver, p_b is the position of the beacon, c is the speed of light and $\Delta\tau$ is a time delay due to phenomena such as processing time in the hardware and unsynchronized clocks and can be determined from calibration experiment, and e_τ is measurement noise. Using (17) the distance between a beacon and a receiver can be inferred from TOA measurements and used in the UWB measurement model (12).

Under *non-line-of-sight* (NLOS) or multipath conditions an extra time delay δ must be inserted into (17) due to the fact that the transmitted signal might not take the shortest path from the beacon to the receiver [37]. The delay δ is unknown and can cause biases in the obtained measurements. It depends on the environment and can change if the environment changes and if the beacon or receiver is moved. This is why it is important to be able to detect NLOS conditions. Due to the high speed of light, a small delay δ or even a small error in the calibration of $\Delta\tau$ will give rise to large positioning errors where the distance between the UWB units seems to be longer than it really is. Attempts have been made to model the distance measurements errors for example in [38, 39], where error probability distributions such as the lognormal and the skew-t distributions are explored. What both these probability density models have in common is long heavy tails for positive errors but shorter tails for negative errors.

In the proposed filter solution the UWB measurement error noise is assumed to be zero mean Gaussian, a distribution with short tails both for positive and negative errors. This means that, in order to get good filtering results, measurements belonging to the long positive tails of a more realistic error distribution must be detected and rejected in the filter. This is here done by the same type of chi-squared test as described in (16). Note that this approach can introduce a small bias in the pose estimates since the actual measurements are not zero mean. It is however a good way to make the filter more robust against NLOS measurements without having to model the errors, which is something that can be difficult in indoor environments prone to change.

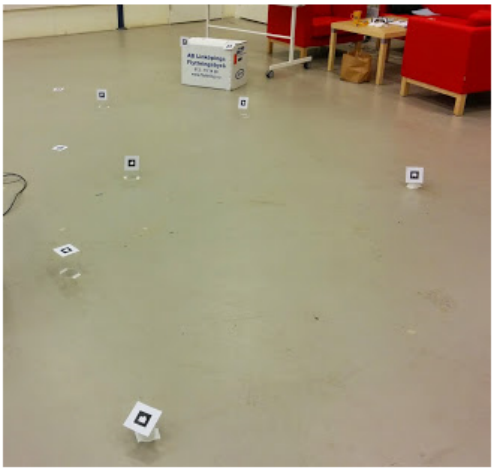


Fig. 4: The scene in which the sensor platform was moved.

IV. Experiments

In this section we describe the experiments, the preprocessing used for synchronization of the sensors, analysis of the UWB data, and estimation results using several sensor combinations and geometrical configurations.

A. Experimental Design

Measurement data was collected from an indoor environment where markers, AR-tags [40], were placed on floor level, see Fig. 4. Five ($n^b = 5$) Spoonphone [41] UWB beacons with unique ID:s were placed around the area, see Fig. 6. The purpose of the markers is to increase the number of possible corners in the images since the floor was otherwise quite featureless. Hence, at no time has any marker-based localization methods been used to simplify the SLAM problem.

Two different sensors were mounted on a rigid platform which is shown in Fig. 5. The first sensor is a Spoonphone with a UWB radio signal receiver application, which recorded the distances and corresponding ID:s to the UWB beacons, and a built in IMU. The Spoonphone UWB system utilizes a two-way-ranging technique which enables computation of $\Delta\tau$ in (17) online. The measurements obtained from this system are hence already compensated for this term even without calibration. The second sensor is an Xsens IMU [42] with an 480×640 pixel gray scale camera contained in a single unit. The sensor platform was hand held while moved over the area with markers on the floor. The IMU recorded measurements at 100 Hz and the camera at 12.5 Hz. The UWB sampling time was not uniform but each beacon normally transmitted a couple of times each second.

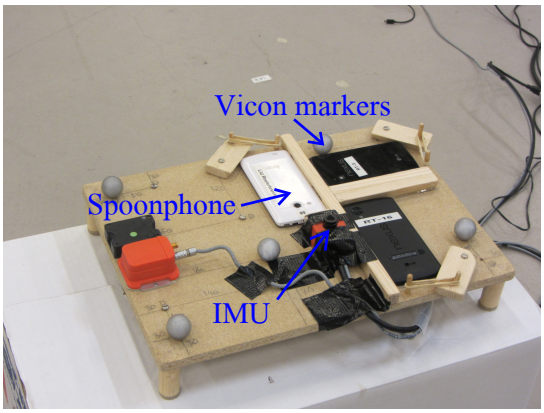


Fig. 5: The sensor platform used for data collection.

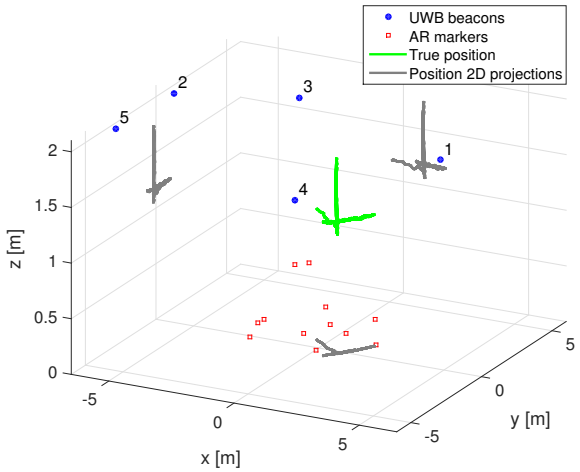
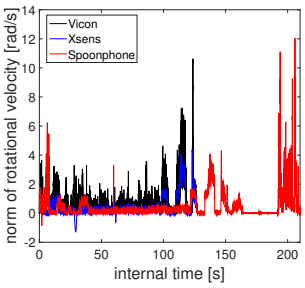
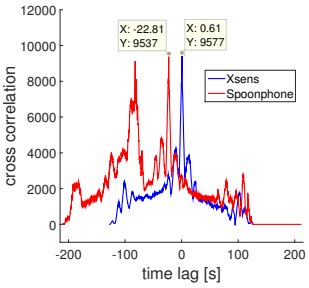


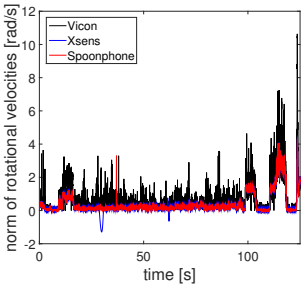
Fig. 6: An illustration of the data collection set up and the motion ground truth.



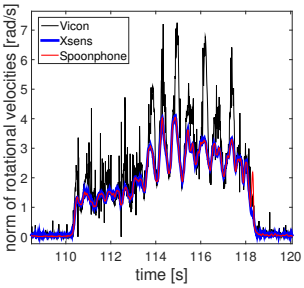
(a) Data before synchronization.



(b) The maximum of the correlation with the Vicon data suggest that the Xsens clock is 0.61 s slower than the Vicon clock and that the Spoonphone clock is 22.81 s faster.



(c) Data after synchronization.



(d) Zoom in of data after synchronization.

Fig. 7: An illustration of the data synchronization. The norm of the rotational velocities recorded with the Vicon, Xsens, and Spoonphone sensors, respectively, is used. Note that interpolation has been done in order to get uniform sample times resulting in shorter periods where the norm is negative. The maximum correlation gives the time lag between the internal clocks.

A Vicon system [43] with sub-centimeter and sub-degree accuracy was used to record ground truth data which is used for filter performance evaluation. The Vicon system was also used to measure the positions of the UWB transmitters, which are considered as known parameters in the filter. An illustration of the data collection setup can be seen in Fig. 6.

The Xsens sensor-unit, Spoonphone, and Vicon system all have their own internal clock and synchronization was done after the data collection. The norm of the rotational velocity of the sensor platform was used in the synchronization process. This norm was computed from the Vicon data, the Xsens gyroscopes and the Spoonphone gyroscopes, respectively. Then interpolation was performed such that the sample time for all the three sensors were the same and also uniform. The three sensors were then synchronized by maximizing the correlation between the processed gyroscope norms. The clock in the Vicon data was used as reference to which the other two sensors were to be synchronized. Each recorded dataset started and ended with a couple of seconds where the sensor platform was vividly shaken to excite the gyroscopes so that the described correlation technique would work well. These parts of the data were however only used in the synchronization process, not for filtering. An example of recorded data before and after synchronization can be seen in Fig. 7.

An analysis of the collected UWB distance measurements was performed. Measurements were compared to the recorded ground truth and error probability histograms for all the recorded data can be seen in Fig. 8. The specified accuracy of the Spoonphone system was about 1 dm. The error histograms however clearly shows that this is not the case here. All histograms have a high probability mode at around 0.5 m, probably representing measurements where the transmitted signals did not take the shortest path due to either multipath or NLOS. The error histogram for UWB beacon number three, and to some degree beacon number five, also has a smaller peak at approximately -1 dm. This smaller peak closer to zero probably represents LOS measurements but could also be measurement error of the true distance or a small clock synchronization error. From the error histograms it is clear that LOS measurements are rare. Also, taking a closer look at the data revealed that it suffers from periods where the obtained distances are locked on to the exact same value as the previous measurement. In fact, almost 70% of the measurements suffered from this condition. All UWB measurements were therefore discarded and replaced by simulated measurements

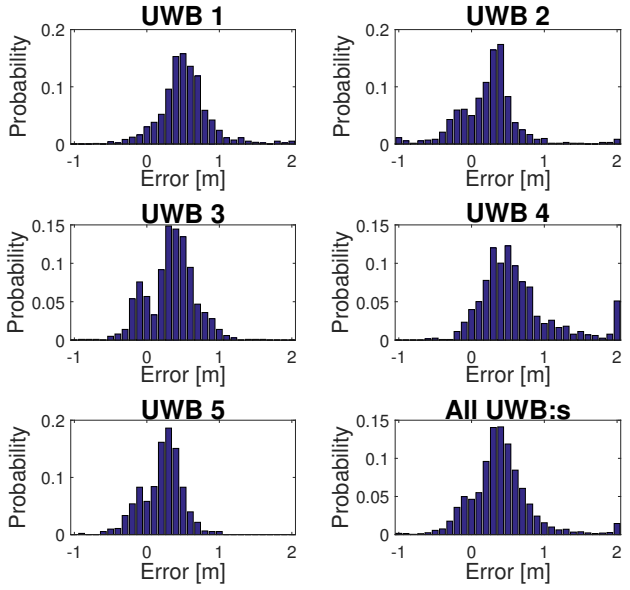


Fig. 8: The illustration shows distance measurement error histograms for the different UWB beacons after removing all measurements that are locked on to the exact same value as the previous measurement.

based on the ground truth data measured with the Vicon system. The sample times of the simulated measurements correspond exactly to the sample times of the real recorded data and the measurements were perturbed with Gaussian noise with standard deviation 1 dm, which is roughly the specified accuracy. Even though this was not the accuracy obtained with the equipment used in this paper it is likely to believe that this accuracy can be obtained with some other equipment. The IMU data and images used for filter evaluation in Section IV-B are however both real data from the experiments.

The dataset used for filter evaluation in this paper is 30 s long. The sensor platform has been moved in all three directions (x, y and z) while being rotated. The acceleration and velocity of the sensor platform is moderate. During four shorter time periods of the data there are interruptions in the flow of UWB measurements. The ground truth trajectory can be seen in Fig. 6.

B. Results

In this section a selection of experimental results obtained by applying the algorithm proposed in Sec. II and Sec. III on the experimental data described in Sec. IV-A are presented.

1) Different sensor configurations: The use of data from the different sensors placed on the sensor platform was turned on or off in order to investigate whether there was any gain in using all three sensors in comparison to only one or two of them. The experimental results are summarized in Table I where the *root mean square errors* (RMSEs) for different sensor combinations are presented.

The upper part of the table shows results obtained when using data from the UWB system but not from the camera. Here we can see that it is possible to obtain relatively good

TABLE I: Estimation RMSE when using different sensor combinations. Distance errors are given in centimeters and angle errors in degrees. Experiments including the camera measurements have been made both with the *direct parametrization* (DP) and the *inverse depth parametrization* (IDP) of the visual landmarks.

Sensors	x	y	z	Roll	Pitch	Yaw
UWB[1]	6.5	7.9	16.7	–	–	–
UWB + Gyro	6.4	8.6	17.3	29.1	26.9	19.3
UWB + IMU	8.0	20.8	21.6	1.0	3.3	5.0
Vision + Gyro, DP	11.8	8.5	10.1	1.6	1.9	2.4
Vision + Gyro, IDP	10.1	8.9	6.7	1.8	2.9	2.8
Vision + IMU, DP	17.6	11.8	12.0	2.0	2.0	1.0
Vision + IMU, IDP	12.2	13.8	5.0	1.3	1.8	1.2
Vision + Gyro + UWB, DP	3.3	3.5	3.6	1.5	1.0	1.5
Vision + Gyro + UWB, IDP	5.5	5.6	5.6	0.9	1.2	2.2
Vision + IMU + UWB, DP	4.5	4.9	3.5	1.2	1.1	1.9
Vision + IMU + UWB, IDP	5.2	6.5	6.3	1.3	1.1	2.1

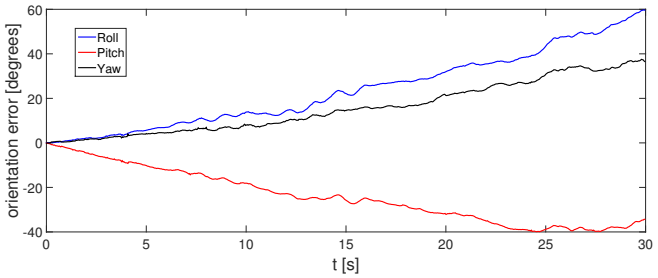


Fig. 9: An illustration of the orientation estimate drift when using only data from UWB and gyroscopes.

position estimates in the x- and y-axis direction but the z-axis direction is more difficult. This was also noticed in [26]. A reasonable explanation is that this problem appears because of the setup of the UWB beacons with a high spread in the x- and y-directions but a very low spread in the z-direction.

As mentioned before, with only UWB data the orientation of the platform cannot be estimated. Adding data from the gyroscopes makes it possible to estimate orientation based on an correct initial orientation but the biased measurements make the estimates drift, as can be seen in Fig. 9, and the UWB data contains no information to reduce this drift. Also adding data from the accelerometers makes it possible to correct the drifting orientation estimates by alignment of linear acceleration information obtained indirectly from the UWB measurements and directly from the accelerometers. However, the position estimate get much less accurate. During the time periods when no UWB measurements were obtained, the filter has to rely only on dead reckoning from biased accelerometer data to estimate the position. During these periods the position estimation error grows quickly. The filter is able to recover when new UWB measurements are obtained but the total RMSE for the whole experiment is affected. Fig. 10 illustrates this behavior.

The middle part of the Table I shows results obtained when using data from the camera but not from the UWB system. Here we can see that the orientation estimates are better compared to in the upper part of the Table I. Fig. 11 illustrates that the visual-inertial combination has a tendency

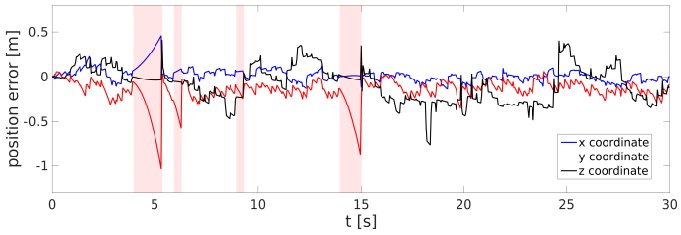


Fig. 10: An illustration of the position estimate drift during periods without any new UWB measurements for the case when using data only from UWB and IMU. The time intervals shaded red indicate periods without UWB measurements.

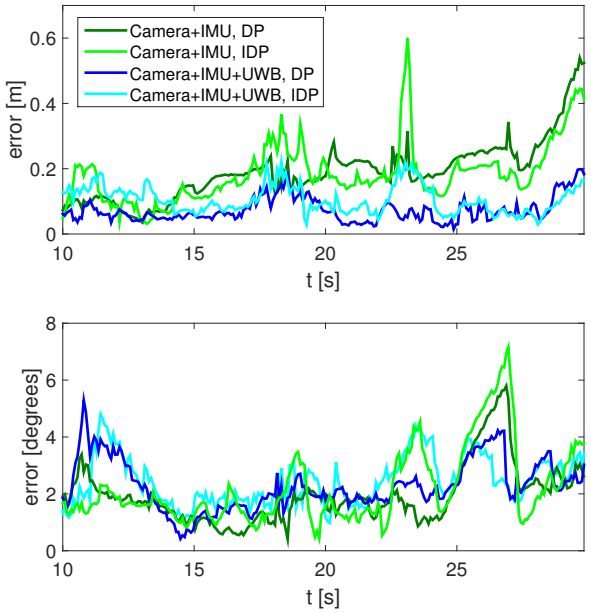


Fig. 11: An illustration of the position errors (top) and orientation errors (bottom) using various sensor combinations.

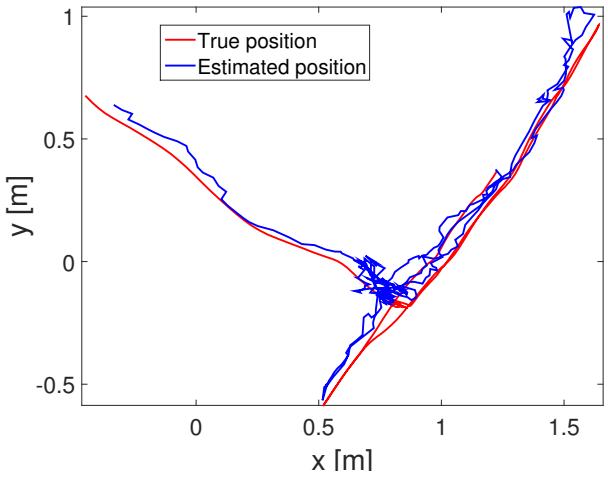


Fig. 12: An illustration of the tracking result when combining data from all three sensors.

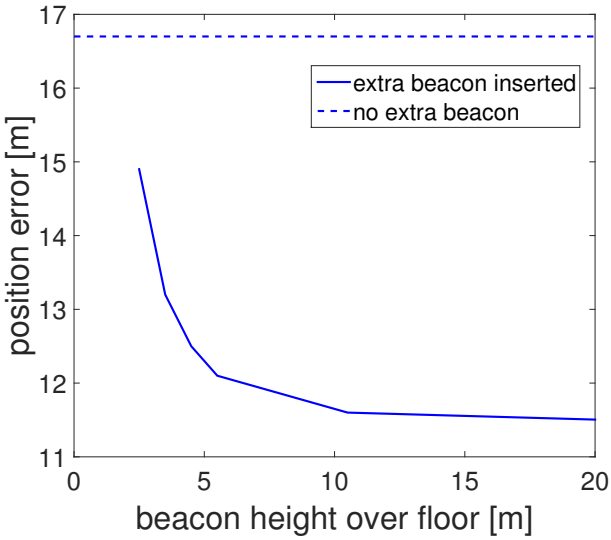


Fig. 13: An illustration of how the positioning accuracy in the z-direction is increased when an additional beacon, centered in the middle of the original beacons but at varying height, is introduced. The dashed line corresponds to the error from Section IV-B1 with the original setup and the solid lines corresponds to the error when the additional beacon is introduced.

to drift, a phenomena that was typical for all four experiments from the middle part of the Table I.

The lower part of Table I table shows the results after combining all three sensors. The table shows clear benefits of combining visual-inertial SLAM with UWB measurements. The over all position and orientation estimates have improved. There is no longer any issue with estimation of position in the z-axis direction nor with maintaining good estimates during the periods where no new UWB measurements are obtained and Fig. 12 shows that drift is no longer present, or at least it has been reduced considerably.

It should be pointed out that all the results obtained without UWB are local in nature, and relies on being properly initialized. Similarly, only UWB and gyroscope can only give a local orientation based on the initialization.

2) Effect of UWB beacon placement: In indoor environments it is common to place UWB beacons at high levels so that the direct path between the beacon and the receiver is less likely to be blocked by *e.g.* furniture or people. This section will therefore only consider beacons placed at high levels.

The UWB system in its current configuration has some problems estimating the height of the sensor platform, as could be seen in Section IV-B1. In order to investigate whether the pose estimate accuracy could be further increased, simulated UWB beacons were placed such that the spread in height became larger. In a first try, one additional beacon centered in the middle of all the other beacons was simulated. The height at which it was placed was varied and the tracking result using only UWB data can be seen in Fig. 13.

In a second experiment the additional beacon was instead

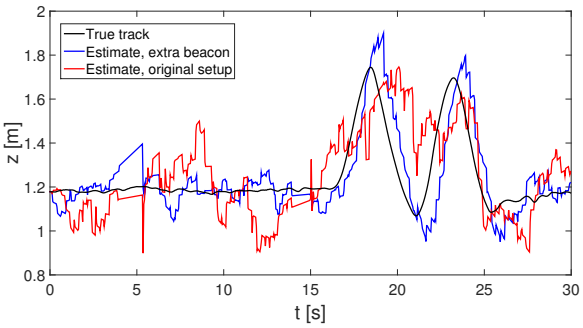


Fig. 14: An illustration of the tracking result in the z-axis direction with and without an extra simulated beacon placed at roof level (2.5 m) right above the sensor platform track.

placed right above the center of the track at the reasonable roof level 2.5 m. This resulted in a tracking accuracy in z-direction of about 11.5 cm, which is the lower limit in Fig. 13. Raising the beacon higher gave no significant improvement. Neither did adding more than one beacon at the same height 2.5 m.

These results tells us that spreading the UWB beacons in the z-axis as well gives a better geometrical configuration for the case when only UWB data is used which should also be clear from Fig. 14. The highest accuracy is achieved when an extra tag is placed right above the track or alternatively at a very high level. However, even though there is a clear advantage with adding one extra beacon for the case when only UWB data is used in the filter, turning on all three sensors gave no clear improvement compared to the last part of Table I. This indicates that the camera-inertial system can support the UWB system in dimensions where the UWB beacon configuration fails to give sufficient information. This raises the questions; Can any of the UWB beacons in the original setup be removed without great loss in tracking accuracy? How few UWB beacons are actually needed for the case when data from camera, IMU and UWB are all being fused?

3) Minimal UWB beacon setup: This section evaluates how many and which beacons are needed to improve the visual-inertial SLAM solution. This is done by removing one beacon (from the original setup in Fig. 6) at a time from the measurements in the "Vision + IMU + UWB, DP" configuration in Sec. IV-B1. In each stage, the beacon that changes the estimate accuracy the least compared to the previous best setup is removed. The result of the procedure, is presented in Table II. It shows that the difference between using five or two beacons is almost negligible, yielding a position and orientation accuracy of about half a decimeter and a few degrees respectively. With less than two beacons, the estimate degenerates considerably, which is in agreement with the observability discussion in Sec. II-F

TABLE II: RMSE using different subsets of UWB beacons from the configuration in Fig. 6 in the "Vision + IMU + UWB, DP" configuration in Sec. IV-B1. Distance and orientation errors are given in centimeters and degrees, respectively. Lines in bold face, indicate the beacons contributing the least to the estimate for a given number of beacons.

ID:s of removed UWB beacons	x	y	z	Roll	Pitch	Yaw
–	4.5	4.9	3.5	1.2	1.1	1.9
1	**4.6**	**4.5**	**3.6**	**1.5**	**1.2**	**2.0**
2	4.5	4.3	4.5	0.9	1.0	2.3
3	4.7	5.0	3.5	1.5	1.1	2.1
4	4.6	4.6	4.5	1.0	1.1	2.3
5	4.9	5.5	3.6	1.4	1.0	1.9
1,2	**5.0**	**4.4**	**4.3**	**1.6**	**1.1**	**2.0**
1,3	4.9	5.1	4.0	1.8	1.4	2.0
1,4	4.8	4.9	4.5	1.1	1.2	2.3
1,5	5.8	5.7	3.9	1.3	1.4	1.6
1,2,3	**5.7**	**4.8**	**4.2**	**1.8**	**1.2**	**2.1**
1,2,4	8.9	5.1	6.1	1.5	1.2	2.7
1,2,5	7.9	7.0	6.8	1.7	1.5	1.4
1,2,3,4[2]	–	–	–	–	–	–
1,2,3,4	**11.8**	**7.1**	**9.0**	**1.4**	**1.5**	**1.6**

V. Conclusions

Indoor positioning systems based on measurements from cameras and/or inertial measurement units are gaining popularity. Systems like this often suffer from drifting estimates since both these sensors measure only relative motion. Existing solutions to this problem predominantly rely on pre-installed infrastructure, such as reference markers of some type, or detailed maps. Additional infrastructure is both time consuming and costly to install and this is a limiting factor for the applicability of visual-inertial positioning systems.

The use of time of arrival measurements from ultra wide band to a few number of anchor points in order to support visual-inertial positioning systems has been explored in this paper. It has been shown that measurements of the distance to a few anchor points using UWB can be used to aid visual-inertial simultaneous localization and mapping to obtain improved drift-free global six degree of freedom pose estimates. It has been shown that the proposed solution is capable of handling missing UWB data and also poorly placed anchor points. Furthermore, this can be achieved without the need for extensive pre-mapping of the environment since the results presented in this paper has shown that only a very few number of anchor points is needed. The result has been verified with experimental data with good ground truth.

Future work should include exploring more sophisticated visual features that enable appearance based loop closures for robuster performance. It is also important to further investigate if the Gaussian distribution assumption of the UWB distance measurement errors is reasonable. Based on the results in this paper this might not be the case since non-line-of-sight measurements seems to occur rather frequently. If non-Gaussian assumptions has to be made, then perhaps a particle filter implementation of the proposed filter would give better results. A third direction for future work is to also treat the positions of the UWB anchor points as landmarks within the SLAM framework.

Acknowledgment

This paper was funded by the Swedish Foundation for Strategic Research through the grants VPS (IIS11-0081) and CoopLoc, by the Swedish Research Council through the Linnaeus Environment CADICS, and by the Swedish strategic research center Security Link.

High accuracy reference measurements are provided through the use of the Vicon real-time tracking system courtesy

of the UAS Technologies Lab, Artificial Intelligence and Integrated Computer Systems Division (AIICS) at the Department of Computer and Information Science (IDA), Linköping University, Sweden[3].

REFERENCES

[1] G. A. Thomas, J. Jin, T. Niblett, and C. Urquhart, "A versatile camera position measurement system for virtual reality TV production," in *International Broadcasting Convention*, Sep. 1997, pp. 284–289.

[2] T. De Gaspari, A. C. Sementille, D. Z. Vielmas, I. A. Aguilar, and J. F. Marar, "ARSTUDIO: A virtual studio system with augmented reality features," in *Proceedings of the 13th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, 2014, pp. 17–25.

[3] Y. Yokokohji, Y. Sugawara, and T. Yoshikawa, "Accurate image overlay on video see-through hmds using vision and accelerometers," in *Proceedings of the IEEE Virtual Reality 2000 Conference*, 2000, pp. 247–.

[4] TotalTrax Inc. (2015, Apr.) TotalTrax, Inc. Track. Measure. Manage. [Online]. Available: http://www.totaltraxinc.com/

[5] A. Mulloni, D. Wagner, I. Barakonyi, and D. Schmalstieg, "Indoor positioning and navigation with camera phones," *IEEE Pervasive Computing*, vol. 8, no. 2, pp. 22–31, Apr. 2009.

[6] S. Lee and J.-B. Song, "Mobile robot localization using infrared light reflecting landmarks," in *International Conference on Control, Automation and Systems*, Oct. 2007, pp. 674–677.

[7] F. Bonin-Font, A. Ortiz, and G. Oliver, "Visual navigation for mobile robots: A survey," *Journal of intelligent and robotic systems*, vol. 53, no. 3, pp. 263–296, 2008.

[8] D. H. Titterton and J. L. Weston, *Strapdown Inertial Navigation Technology*, ser. IEE Radar, Sonar, Navigation and Avionics. Stevenage, UK: Peter Peregrinus Ltd., 1997.

[9] S. Rady, A. Kandil, and E. Badreddin, "A hybrid localization approach for UAV in GPS denied areas," in *IEEE/SICE International Symposium on System Integration*, Dec. 2011, pp. 1269–1274.

[10] E. S. Jones and S. Soatto, "Visual-inertial navigation, mapping and localization: A scalable real-time causal approach," *International Journal of Robotics Research*, vol. 30, no. 4, pp. 407–430, 2011.

[11] E. Foxlin and L. Naimark, "VIS-Tracker: A wearable vision-inertial self-tracker," *Virtual Reality*, vol. 3, p. 199, 2003.

[12] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, Nice, France, Oct. 13–16 2003, pp. 1403–1410.

[13] E. Eade, "Monocular simultaneous localisation and mapping," Ph.D. dissertation, Cambridge University, 2008.

[14] A. Kim, , and R. M. Eustice, "Real-time visual SLAM for autonomous underwater hull inspection using visual saliency," *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 719–733, Jun. 2013.

[15] P. Pinies, T. Lupton, S. Sukkarieh, and J. Tardos, "Inertial aiding of inverse depth SLAM using a monocular camera," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Roma, Italy, 10–14 Apr. 2007, pp. 2797–2802.

[16] Z. Sjanic, M. A. Skoglund, F. Gustafsson, and T. B. Schön, "A nonlinear least squares approach to the SLAM problem," in *Proceedings of the IFAC World Congress*, vol. 18, Milan, Italy, 28–2 Aug./Sep. 2011.

[17] M. Li and A. I. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *International Journal of Robotics Research*, vol. 32, no. 6, pp. 690–711, May 2013. [Online]. Available: http://dx.doi.org/10.1177/0278364913481251

[18] M. Bryson, M. Johnson-Roberson, and S. Sukkarieh, "Airborne smoothing and mapping using vision and inertial sensors," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, 2009, pp. 3143–3148.

[19] R. Jiang, R. Klette, and S. Wang, "Modeling of unbounded long-range drift in visual odometry," in *Image and Video Technology (PSIVT), 2010 Fourth Pacific-Rim Symposium on*, Nov. 2010, pp. 121–126.

[20] J. Chandaria, G. A. Thomas, and D. Stricker, "The MATRIS project: real-time markerless camera tracking for augmented reality and broadcast applications," *Journal of Real-Time Image Processing*, vol. 2, no. 2-3, pp. 69–79, 2007.

[21] D. Stojanović and N. Stojanović, "Indoor localization and tracking: Methods, technologies and research challenges," *Facta Universitatis, Series: Automatic Control and Robotics*, vol. 13, no. 1, pp. 57–72, 2014.

[22] J. Mehta and C. R. Varnagar, "A review and insight on various indoor positioning and navigation techniques," in *National Conference on Emerging Trends in Computer, Electrical and Electronics (ETCEE-2015)*, Rajkot, India, 14 Mar. 2015.

[23] C. Zhang, M. Kuhn, B. Merkl, A. Fathy, and M. Mahfouz, "Real-time noncoherent UWB positioning radar with millimeter range accuracy: Theory and experiment," *IEEE Transactions on Microwave Theory and Techniques*, vol. 58, no. 1, pp. 9–20, Jan. 2010.

[24] S. Sczyslo, J. Schroeder, S. Galler, and T. Kaiser, "Hybrid localization using UWB and inertial sensors," in *IEEE International Conference on Ultra-Wideband*, vol. 3, Sep. 2008, pp. 89–92.

[25] S. Pittet, V. Renaudin, and B. Merminod, "UWB and MEMS based indoor navigation," *Royal Institute of Navigation*, vol. 61, no. 3, pp. 369–384, 2008.

[26] J. D. Hol, F. Dijkstra, H. Luinge, and T. B. Schon, "Tightly coupled UWB/IMU pose estimation," in *IEEE International Conference on Ultra-Wideband*, Sep. 2009, pp. 688–692.

[27] A. Benini, A. Mancini, and S. Longhi, "An IMU/UWB/vision-based extended Kalman filter for mini-UAV localization in indoor environment using 802.15. 4a wireless sensor network," *Journal of Intelligent & Robotic Systems*, vol. 70, no. 1-4, pp. 461–476, 2013.

[28] Z. Sjanic, M. A. Skoglund, T. B. Schön, and F. Gustafsson, "A nonlinear least-squares approach to the SLAM problem," in *Proceedings of 18th IFAC World Congress*, Milan, Italy, 28–2 Aug./Sep. 2011.

[29] J.-Y. Bouguet, "Camera calibration toolbox for Matlab," http://www.vision.caltech.edu/bouguetj/calib_doc/, 2010.

[30] J. Hol, T. B. Schön, and F. Gustafsson, "Modeling and calibration of inertial and vision sensors," *The International Journal of Robotics Research*, vol. 29, no. 2, Feb. 2010.

[31] J. M. M. Montiel and A. J. Davison, "A visual compass based on SLAM," in *Proceedings the IEEE International Conference on Robotics and Automation (ICRA)*, Orlando, Florida, USA, May 2006, pp. 1917–1922.

[32] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *Autonomous robot vehicles*. New York, NY, USA: Springer-Verlag New York, Inc., 1990, pp. 167–193.

[33] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Conference*, Manchester, UK, 1988, pp. 147–151.

[34] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh International Conference on Computer Vision (ICCV'99)*, Corfu, Greece, 1999, pp. 1150–1157.

[35] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *International Journal of Robotics Research*, 2010.

[36] D. Dardari, A. Conti, U. Ferner, A. Giorgetti, and M. Z. Win, "Ranging with ultrawide bandwidth signals in multipath environments," *Proceedings of the IEEE*, vol. 97, no. 2, pp. 404–426, Feb. 2009.

[37] J. D. Hol, "Sensor fusion and calibration of inertial sensors, vision, ultra-wideband and GPS," Dissertations No 1368, Linköping Studies in Science and Technology, Jun. 2011.

[38] N. Alsindi, B. Alavi, and K. Pahlavan, "Measurement and modeling of ultrawideband TOA-based ranging in indoor multipath environments," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 3, pp. 1046–1058, Mar. 2009.

[39] H. Nurminen, T. Ardeshiri, R. Piche, and F. Gustafsson, "Robust inference for state-space models with skewed measurement noise," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1898–1902, Nov. 2015.

[40] M. Fiala, "ARTag, a fiducial marker system using digital techniques," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, Jun. 2005, pp. 590–596.

[41] BESPOON. (2015, Mar.) BeSpoon we position. [Online]. Available: http://spoonphone.com/en/

[42] Xsens. (2015, Mar.) Xsens — the leading innovator in 3d motion tracking technology. [Online]. Available: https://www.xsens.com/

[43] Vicon Motion Systems Ltd. (2015, Apr.) Vicon. [Online]. Available: http://www.vicon.com

[3] http://www.ida.liu.se/divisions/aiics/aiicssite/index.en.shtml