Abstract—We consider an indoor tracking system consisting of an inertial measurement unit (IMU) and a camera that detects markers in the environment. There are many camera based tracking systems described in literature and available commercially, and a few of them also has support from IMU. These are based on the best-effort principle, where the performance varies depending on the situation. In contrast to this, we start with a specification of the system performance, and the design is based on an information theoretic approach, where specific user scenarios are defined. Precise models for the camera and IMU are derived for a fusion filter, and the theoretical Cramér-Rao lower bound and the Kalman filter performance are evaluated. In this study, we focus on examining the camera quality versus the marker density needed to get at least a one mm and one degree accuracy in tracking performance.

I. INTRODUCTION

There are today three kinds of tracking systems for indoor use:

- Indoor positioning systems with the purpose to find the approximate position inside a building. There are foot-mounted inertial measurement units (IMU) for rescue personnel [13], as well as smartphone applications based on radio fingerprinting, possibly in combination with sensor fusion with IMU measurements [8][18]. The accuracy of such systems is in the order of meters, and the coverage is a whole or large parts of a building.

- Indoor reference tracking systems, where the VICON system [17] is state of the art. The accuracy is in the order of one mm for position and one degree for orientation, while the coverage is restricted to a part of a room.

- Computer control tracking systems such as game controls and pointing devices (mice, touch screens, joysticks, etc). The accuracy of game controls such as Nintento Wii, Xbox Kinect and Playstation Move is in the order of a decimeter for position and several degrees for orientation, and the coverage is a few square meters [10][12].

Our goal is to investigate feasibility of a tracking system, consisting of consumer grade sensors, that combines the properties of all these systems. The system should have an (i) accuracy comparable to a reference system, an (ii) operating range that covers several rooms, a (iii) cost that matches game controls, and (iv) being easy to deploy and use. More specifically, the combination of an IMU with vision information is investigated. Basically, the system should be possible to run on a smart-phone.

Dead-reckoning IMU measurements provides linear drift in orientation and cubic drift in position. The idea is that opportunistic or pre-deployed markers in the room are detected by the vision system, and used to stabilize the drift [14][1][11][5]. The world coordinates of these markers are assumed known to the system, either from the deployment or a dedicated calibration experiment [7]. The projection of the world coordinates onto the image plane provides constraints to the motion, that can be used to estimate the biases in the IMU sensors and eliminate the drift.

This combination of vision with IMU has been used in several studies for augmented reality (AR), see for instance [2][6]. However, the application of AR requires high accuracy of position and orientation (pose) projected onto the image plane, and the absolute pose accuracy is not the main design issue. AR is essentially a 3 DoF application where orientation angles pan (yaw), tilt (pitch) and roll are most important. Depth is not that sensitive, and small lateral and vertical movements are hard to distinguish from small tilt and pan motions.

The motivation to this study comes from a project called Virtual Photo Set (VPS). The VPS approach is an extension to what is called image based lighting (IBL) [3], in which a panoramic high dynamic range (HDR) image captured in the real scene is used as illumination information during rendering. A key limitation of traditional IBL is that it approximates the illumination as a single spherical distribution of incoming light, captured at a single point in the real scene. To include also important spatial variations in the illumination, a VPS is captured using HDR-video instead of a still image. It consists both of a reconstructed geometric model of the scene and accurate radiometric information describing the intensity and color of light sources as well as the appearance of the surfaces and materials in the scene [16]. Figure 1 illustrates the VPS concept and the realism attainable in the renderings. Reconstruction of detailed VPS models requires accurate tracking of the HDR-camera during capture. Previously, a mechanical tracking system has been used, which suffers from a limited coverage and a difficulty to move to other buildings.
Fig. 1. Illustration of the VPS application. a) The real environment, a photo studio. b) The VPS model consist of a recovered geometric model of the scene that is textured with the photometric information from the HDR-video sequences, and describes how the illumination varies between different locations in the scene. c) Virtual furnitures placed in the recovered VPS model. d) Photo realistic rendering of the virtual furnitures.

The performance in terms of absolute accuracy of pose will be studied theoretically using both simulations and Cramér-Rao analysis. Aspects that affect the performance include the density of the visual landmarks, the quality of the camera, and how much excitation is needed from the sensor platform trajectory.

The outline is as follows. Section II introduces the different coordinate systems needed in the models. Section III and IV describe how the IMU and the visual camera can be modeled. An Extended Kalman Filter (EKF) has been used to solve the tracking problem and this is described in Section V. Numerical evaluation results are presented in Section VI while the results from a CRLB analysis can be seen in Section VII. Last section contains some remarks on the results and a brief discussion of how to proceed with this work.

II. COORDINATE SYSTEMS

This section introduces the different coordinate systems, and describes how transformations between them can be performed on a principal level.

A. Inertial system - I

The inertial system, has its origin in the center of the earth. Its axes are fix relative to fix distant stars. Hence, the system is not spinning along with the earth which means that phenomena such as the Coriolis effect can be taken into account.

B. Earth fix, ECEF - E

The earth centered earth fixed (ECEF) system spins, in contrary to the inertial system, along with the earth. Its z-axis points towards the north pole and its x-axis points towards the crossing of the equator and the Greenwich meridian line.

C. Room fix - R

The room fix coordinate system has its origin in the geodetic latitude and longitude coordinates $\mu$ and $\iota$ and it axes form a NED (North East Down) system. A transformation matrix $T_{RE}$ from ECEF to NED can for example be computed from the $\mu$ and $\iota$ coordinates.

D. Body fix - B

The body fix coordinate system is illustrated in Figure 2. It is fixed in the IMU unit and its axes coincide with the three axes along which the IMU measurements are made. Transformations between the room fix and the body fix systems can for example be represented by three rotation angles $\phi_{BR}, \theta_{BR}$ and $\psi_{BR}$ (x-y-z-conversion) or quaternions. The quaternion representation has the advantage of being continuous under integration and does not suffer from gimbal locks like the angle representation does. It is therefore used in all computations in this work. Results are however presented as rotation angles since they are more intuitively interpreted.

E. Camera fix - C

The coordinate system fixed in the camera has its origin in the camera lens as shown in Figure 2. Its x-axis points away from the camera sensor along the optical axis of the camera. The z-axis is aligned according to the camera sensor in such a way that it points downwards along the sensor. A transformation between the body fix (IMU fix) and the camera fix system can be obtained from calibration experiments [7].

Fig. 2. An illustration of the body fix and the camera fix coordinate systems. These two systems are fix, but possibly rotated, relative to each other.

III. INERTIAL SENSOR MODEL

The inertial measurement unit (IMU) measures acceleration $\mathbf{a}_{BI}$ and angular velocity $\mathbf{w}_{BI}$ of the IMU relative to the inertial system. The measurements are made along three orthogonal axes of the IMU, which for this problem formulation coincide with the body fix coordinate system (see Section II-D). Data from the IMU could suffer from measurement errors due to several reasons. Errors which the IMU can be calibrated for are for example scale errors, non-linearities, cross-talk between the three channels along which the measurements are taken and g-sensitivity. Two type of errors that can not be compensated

\footnote{The gravitation can have influence not only in the acceleration measurements but also in the angular velocity measurements. This is referred to as g-sensitivity}
for are zero mean white measurement noise, $\epsilon_a^\alpha$ and $\epsilon_\omega^\alpha$ and slowly changing biases, $\delta_a^\alpha$ and $\delta_\omega^\alpha$. These two errors have great impact on the measurement accuracy and must therefore be included in the IMU model. The IMU model can then be written down as

$$\pi_a = \epsilon_{BI}^{\alpha} + \delta_a^\alpha + \epsilon_a^\alpha$$  \hspace{0.5cm} (1)$$

$$\pi_\omega = \epsilon_{BI}^\omega + \delta_\omega^\omega + \epsilon_\omega^\omega$$  \hspace{0.5cm} (2)$$

The biases $\delta_a$ and $\delta_\omega$ can in turn be modeled simply as random walks according to

$$\dot{\delta}_i = \eta_i, \; i \in \{\alpha, \omega\}$$  \hspace{0.5cm} (3)$$

where $\eta_i$ is zero mean white noise. Rewriting (1) and (2) taking into account the different coordinate frames, of which some move relative to each other, gives\(^2\)

$$\pi_a = T_{BR} (\pi_{BR}^R - \eta^R) + \delta_a^\alpha + \epsilon_a^\alpha$$  \hspace{0.5cm} (4)$$

$$\pi_\omega = T_{BR} T_{RE} \omega_{EI}^E + \delta_\omega^\omega + \epsilon_\omega^\omega$$  \hspace{0.5cm} (5)$$

where $\eta$ denotes the gravity and $\omega_{EI}$ denotes the spin of the earth. Here parts of interest for the tracking is marked by red. $T_{BR}$ and $T_{RE}$ are transformation matrices and how these can be computed has been described in Section II.

### IV. Camera Sensor Model

Landmarks placed in the tracking environment can be captured by a camera and the image coordinates of the landmarks can be used as measurements of the equipment pose. A suitable model for the 3D to 2D projection of world coordinates onto the camera sensor has to be known if these measurements are to be useful. The observation model used in this work is based on the thick lens model. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

camera lens may therefore not be that easily neglected. The reason for this is that a modern camera is built out of lens systems with several lenses put in sequence and the thickness of a

...
coordinate)

\[
\begin{align*}
\hat{y}_{\text{proj},1} &= \frac{dy_{\text{proj},\text{focus}}}{d_{\text{focus}}} + a + (d - d_{\text{focus}}) \\
\hat{y}_{\text{proj},2} &= \frac{dy_{\text{proj},\text{focus}}}{d_{\text{focus}}} - (d - d_{\text{focus}})
\end{align*}
\]  
(8)

where \(a\) is the maximum distance from the center of line at the second plane of refraction for a ray entering into the camera. The center of the blurry area can then be computed as,

\[
\begin{align*}
\hat{z}_{\text{proj},1} &= \frac{1}{2} \left( \hat{z}_{\text{proj},1} + \hat{z}_{\text{proj},2} \right) = -\frac{d}{x_{\text{lm},c}} \hat{y}_{\text{lm},c} \\
\hat{z}_{\text{proj},2} &= \frac{1}{2} \left( \hat{z}_{\text{proj},1} + \hat{z}_{\text{proj},2} \right) = -\frac{d}{x_{\text{lm},c}} z_{\text{lm},c}
\end{align*}
\]  
(9)

where also (7) has been used. This is the sought 3D to 2D camera projection. Note that for \(d = d_{\text{focus}}\) this reduces to the projection obtained under the assumption of a sharp image.

V. EXTENDED KALMAN FILTERING

The Extended Kalman Filter (EKF) solves the problem of estimating the state vector \(\mathbf{x}\) of some system given some measurements \(\mathbf{y}\) where

\[
\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{u}, \mathbf{v})
\]  
(10)

\[
\mathbf{y} = h(\mathbf{x}, \mathbf{u}, \mathbf{v})
\]  
(11)

and \(\mathbf{u}\) are some measurable input signal to the system, \(\mathbf{v}\) is process noise and \(\mathbf{y}\) denotes measurement noise. Hence, the motion model \(f(\cdot)\) and the observation model \(h(\cdot)\) for the specific problem in this work need to be formulated.

A. Motion model

First the states \(\mathbf{x}\) has to be chosen. Since the aim is to find the position and orientation of the measurement equipment relative to the room these are obvious choices. Since the motion of the equipment can be arbitrary, quaternions are a better option for describing the orientation than rotation angles representation. If also the velocity of the measurement equipment relative to the room is chosen as states then it can be seen in (18)-(19) that the IMU measurements can be used as inputs to the system. Hence, these are also chosen as states. The IMU biases \(\delta_a\) and \(\delta_\omega\) must of course also be chosen as states. Otherwise the IMU measurements can not be adjusted for these biases and wrong information about the acceleration and angular velocity will hence be used in the filtering. To summarize, the following are the states of the sensor platform.

\[
\begin{align*}
\mathbf{x}_{1:3} &= \mathbf{v}_{BR}^R \\
\mathbf{x}_{4:6} &= \mathbf{v}_{BR}^R \\
\mathbf{x}_{7:10} &= \mathbf{\omega}_{BR}^R \\
\mathbf{x}_{11:13} &= \delta_a \\
\mathbf{x}_{14:16} &= \delta_\omega
\end{align*}
\]  
(12-16)

Now note that taking the time derivatives of the states \(\mathbf{x}_{4:6}\) and \(\mathbf{x}_{7:10}\) results in acceleration and angular velocities for the sensor platform relative to the room respectively. The IMU measurements (4) and (5) can then be used to obtain the following motion model.

\[
\begin{align*}
\mathbf{x}_{1:3} &= \mathbf{v}_{BR}^R = \mathbf{v}_{BR}^R \\
\mathbf{x}_{4:6} &= \mathbf{\omega}_{BR}^R = T_{BR}^R(\mathbf{\omega}_{BR}^R) \left( \mathbf{u}_{a} - \mathbf{\omega}_{BR}^R \mathbf{x}_{7:10} \right) + \mathbf{\omega}_{BR}^R \\
\mathbf{x}_{7:10} &= T_{BR}^R(\mathbf{x}_{7:10}) \left( \mathbf{u}_{a} - \mathbf{\omega}_{BR}^R \mathbf{x}_{7:10} \right) + \mathbf{\omega}_{BR}^R
\end{align*}
\]  
(17-18)

\[
\begin{align*}
\mathbf{x}_{11:13} &= \delta_a \\
\mathbf{x}_{14:16} &= \delta_\omega
\end{align*}
\]  
(19-21)

1) Model parameters: The noise terms \(\mathbf{v}_{a}, \mathbf{\omega}_{a}, \mathbf{\omega}_{a}, \mathbf{\omega}_{a}\) can be modeled as Gaussian noise with zero mean. The covariance of the noise together with the parameters \(\tau_a\) and \(\tau_\omega\) in (20)-(21) should be chosen to fit well to real IMU data. They can be determined for example with Allan variance calibration [4]. There are also some extra parameters aside from these IMU parameters which must be known if the above model are to be useful. These extra parameters are

- \(\mathbf{g}^R\) - the gravity vector measured in the room fix coordinate system. It is reasonable to assume that the gravity vector is

\[
\mathbf{g}^R = \begin{bmatrix} 0 & 0 & g \end{bmatrix}^T
\]  
(22)

- \(T_{RE}\) - the transformation matrix from earth fixed coordinates to room fix coordinates. How this can be determined was discussed in Section II-C.

- \(\mathbf{\omega}_{E}\) - the rotation of the earth measured in the earth fix coordinate system. The earth rotates with approximately \(\mathbf{\omega}_{E} \approx 0.073\) mrad/s. Depending on the accuracy of the IMU this magnitude will be more or less hidden in the measurement noise. It is rather pointless to take the rotation of the earth into account if the IMU is not very accurate. However if the performance of the IMU is better then it could be an idea to estimate the rotation vector of the earth for example as

\[
\mathbf{\omega}_{E} = \begin{bmatrix} 0 & 0 & \mathbf{\omega}_{E} \end{bmatrix}^T
\]  
(23)

B. Observation model

Landmarks with known coordinates are assumed placed out in the tracking environment. A camera can capture images of the environment in which these known landmarks can be found. The 2D image coordinates of a landmark then corresponds to the projection given by (9) derived in the Section IV. These image coordinates will be the measurements upon which the EKF is based. Several landmarks might be captured in the same image which means that the observation model will consist of several pairs of image coordinates, one
pair for each landmark captured in the image. Also, the model contains white zero mean measurement noise. The observation model can be formulated as

$$\mathbf{y} = \begin{bmatrix} y_{\text{proj},1} \\ \vdots \\ y_{\text{proj},n_{lm}} \\ \hat{z}_{\text{proj},1} \\ \vdots \\ \hat{z}_{\text{proj},n_{lm}} \end{bmatrix} + \mathbf{v}_{\text{proj}}$$  \hspace{1cm} (24)$$

where $n_{lm}$ is the number of landmarks captured in the image. At least three landmarks are needed since both the position and the orientation (6 DoF) of the sensor platform are to be tracked.

However, the derivation of the 3D to 2D camera projection led to an expression where $y_{\text{proj}}$ and $\hat{z}_{\text{proj}}$ are expressed in terms of the landmark coordinates relative to the camera fix coordinate system. To fit into the EKF framework, these expressions have to be rewritten in terms of the chosen states. The landmark coordinates $\mathbf{r}_{lm,c}$ can be expressed as

$$\mathbf{r}_{lm,c} = [x_{lm,c} \ y_{lm,c} \ z_{lm,c}]^T = \mathbf{r}_{lm,R} - \mathbf{r}_{BR} - \mathbf{r}_{CB}$$

$$= T_{CB}T_{BR}(\mathbf{r}_{BR}) (\mathbf{r}_{lm,R} - \mathbf{r}_{BR}) - \mathbf{r}_{CB}$$

where $i \in 1, 2, \ldots \ n_{lm}$. Using this together with the camera projection formula results in an observation model which can be used in the EKF framework.

1) Model parameters: The above model contains some parameters which need to be known before the model can be used for tracking. First, the positions $\mathbf{r}_{lm,R}$ of the landmarks in the tracking environment have to be known, as mentioned before. Also the position, $\mathbf{r}_{CB}$ and rotation $T_{CB}$ of the camera fix coordinate system relative to the body fix system must be known. These can be obtained with high accuracy from IMU-camera calibration experiments [7][9]. The camera parameters $t$ and $d$ corresponding to the thickness of the lens and the effective focal length must also be known and should be chosen to fit real data from the camera.

VI. NUMERICAL EVALUATION

The performance of the tracking was evaluated by simulations. Ground truth data together with IMU and camera measurements are simulated in a first step. Then, these simulated measurements are used in an EKF filter with the above motion and observation models and the tracking result is compared with the simulated ground truth.

The IMU model parameters has been taken from the Xsens MTi-10 series datasheet [21]. A 20 mm lens was used as a camera model. The field of view of the camera has not been taken into account explicitly. The simulation scenario has however been constructed in a realistic way for a camera with 85 degrees field of view in the horizontal direction and 60 degrees in the vertical direction. Table I shows how far from the optical axis a landmark can be before it is outside the field of view for some different distances between the camera and the virtual wall at which the landmarks were placed. a) For the first configuration, the distance between two adjacent landmarks is 1 meter. b) The second configuration resembles the first one but the landmark pairs 7-10 and 8-9 have been moved 1m further apart while the pairs 11-14 and 12-13 have been moved 2m further apart.

<table>
<thead>
<tr>
<th>Distance to wall [m]</th>
<th>Horizontal distance from optical axis [m]</th>
<th>Vertical distance from optical axis [m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.4</td>
<td>0.57</td>
</tr>
<tr>
<td>3</td>
<td>2.7</td>
<td>1.7</td>
</tr>
<tr>
<td>5</td>
<td>4.5</td>
<td>2.9</td>
</tr>
<tr>
<td>7</td>
<td>6.4</td>
<td>4.0</td>
</tr>
</tbody>
</table>

A. No motion, varying observation noise

Simulations were made where the sensor platform stood completely still. The simulation scenario were such that the camera on the sensor platform pointed towards a wall with three landmarks with known positions. Only the three landmarks denoted by 1-3 in Figure 4 were used. The accuracy of the camera observations was varied to see how it influenced the accuracy of the tracking. It is reasonable to assume that image coordinates of a landmark can be determined with
approximately the accuracy of one width/height of a pixel. The observation noise \( \varepsilon_{\text{proj}} \) was therefore modeled as Gaussian noise with zero mean and covariance such that there was roughly 95% probability of a measurement being assigned to the correct pixel. Today’s digital cameras have sensors with pixels with widths/heights in the range from 5 \( \mu \text{m} \) for the best system cameras up to around 40 \( \mu \text{m} \). The different choices of observation noises and the corresponding assumed pixel widths for the camera sensor can be seen in Table II.

<table>
<thead>
<tr>
<th>Pixel width, ( 4\sigma ) [( \mu \text{m} )]</th>
<th>( \sigma ), observation noise [( \mu \text{m} )]</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.25</td>
</tr>
<tr>
<td>10</td>
<td>2.5</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
</tr>
</tbody>
</table>

The simulation result can be seen in Figure 5 where also the distance from the sensor platform to the wall with three landmarks has been varied. It can be seen that a camera with a sensor with small pixel sizes must be used if the requirements on the tracking accuracy are to be satisfied. Hence, for the remaining simulations the standard deviation of the observation noise were fixed to 1.25\( \mu \text{m} \) (corresponding to the camera sensor with the highest resolution). However, the figure also shows that if the sensor platform is far away from the landmarks then it is not even enough to have a camera sensor with good resolution. The tracking requirements on the position estimates are not satisfied even for the best camera sensor used in the simulations.

B. No motion, varying landmark configurations

The influence of the landmark configuration were studied to see if the tracking accuracy could be increased further. The two different landmark configurations in Figure 4 were used. Also the number of landmarks from these configurations were varied from \( n_{1\text{lm}} = 3 \) up to \( n_{1\text{lm}} = 14 \) where, for each simulation, the landmarks denoted by 1-\( n_{1\text{lm}} \) were used. The result for the first configuration can be seen in Figure 6. The tracking accuracy was highly increased for long distances between the sensor platform and the wall when the number of landmarks observed by the camera was increased. However, it can be seen that the the requirements were never satisfied for the \( x_{\text{BR}} \)-coordinate position estimate (vertical position), not even when as many as 14 landmarks were used. It can be seen in Figure 7 that the second landmark configuration, where the landmarks has a larger spread in the vertical direction, seems to be the solution to this problem.

C. Moving platform

The sensor platform was moved along some benchmark tracks to see if motion affected the accuracy in the tracking estimates. The following tracks were simulated.

- Rotation about the body fix x-axis with \( \pm 20^\circ \) with a period time of 2 seconds.

\[
\dot{\phi}_{\text{BR}} = \frac{20\pi}{180} \cos \left( \frac{\pi}{2} t \right)
\]  

(26)
Fig. 7. Width of the 95% confidence interval for position and orientation estimates. The std of the observation noise was fixed to 1.25 µm and the platform was still. The plots show the effect on the tracking accuracy of changing the number of observed landmarks from configuration 2. The red dotted line shows the highest acceptable tracking error.

- Rotation about the body fix y-axis with ±20° with a period time of 2 seconds.
  \[ \dot{\theta}_{BR} = \frac{20\pi}{180} \cos \frac{\pi}{2} t \]  

- Rotation about the body fix z-axis with ±20° with a period time 2 seconds.
  \[ \dot{\psi}_{BR} = \frac{20\pi}{180} \cos \frac{\pi}{2} t \]  

- Circular movement with radius of 1 m and a period time of 5 seconds
  \[ \pi_{BR}^k = \begin{bmatrix} -\left(\frac{2\pi}{5}\right)^2 \cos \frac{2\pi}{5} t \\ -\left(\frac{2\pi}{5}\right) \sin \frac{2\pi}{5} t \\ 0 \end{bmatrix} \]  

All landmarks in the second configuration were used in these simulations and the observation noise was fixed to 1.25 µm. It can be seen in Figure 8 that the different benchmark tracks did not affect the tracking accuracy significantly. Hence, moving the platform did not worsen the accuracy compared to when the platform was standing still.

VII. Cramér-Rao Lower Bound

We here study the Cramér-Rao Lower Bound (CRLB) on the tracking accuracy to get a theoretical lower bound on the position and orientation RMSE performance. The CRLB is given by

\[ P_k|k = E \left[ (\hat{x}_{k|k} - x_k)(\hat{x}_{k|k} - x_k)^T \right] \geq J_k^{-1} \]  

where \( P_k|k \) is the covariance matrix for the unbiased state vector estimate \( \hat{x}_{k|k} \) of the true state \( x_k \) at time \( k \).

We choose to compute the parametric CRLB (parCRLB) rather than the more standard Bayesian posterior CRLB (postCRLB) [15]. Essentially, parCRLB depends on the actual true trajectory, while postCRLB is the average over all \textit{a priori} possible trajectories. Since we want to evaluate the importance of excitation from the movement, parCRLB makes sense. It is also much simpler to compute, where basically the Ricatti equation in the EKF is run, where the linearizations are made with respect to the true trajectory, not the estimated one. The results can be seen in Figure 9. If the CRLB would have turned out be significantly smaller than the accuracy obtained in the numeric evaluation, then it would have implied that some other, non-linear, filter could have better performance than the EKF. The CRLB is however only slightly smaller than the accuracy obtained from the numeric evaluation which is an indication of that the EKF is a good choice. Also, one can see that the CRLB computations supports the conclusion that low observation noise and several landmarks are needed to fulfill the requirements.

VIII. CONCLUSIONS

The CRLB computations are consistent with the numeric evaluations and the accuracy obtained from the two evaluation methods does not differ much. This implies that that the EKF 3The indexing \( k|k \) denotes the estimate at time \( k \) given measurements up to time \( k \) in contrary to \( k|k-1 \) which denotes a prediction at time \( k \) given measurements up to time \( k-1 \).
filter performs close to optimal and there is not much to gain from applying some other, non-linear, filter to the problem. The evaluation shows that the requirements on the tracking accuracy can be satisfied by an EKF based on IMU measurements in combination with camera observation of landmarks with known positions under certain conditions. It requires a camera and an image processing algorithm which allows for the landmark image coordinates to be determined with high accuracy. However, it is reasonable to believe that this is not a very big issue since today there are algorithms which can make use of patterns to obtain sub-pixel accuracy. Also the number of landmarks which can be captured in a single image and their relative positions seems to have a great influence over the tracking accuracy. Quite many landmarks are needed to satisfy the requirements. This means that it will be a rather time consuming work to determine the exact position of every landmark if this system is used in reality. It could therefore be interesting to study how the tracking is affected if the observations consist of many more landmarks than used in this work but where the position of these landmarks are uncertain. Such an approach could perhaps result in a much more user friendly system where the landmarks must not be placed with such great care.

**ACKNOWLEDGMENT**

This paper was funded by the Swedish Foundation for Strategic Research through grant IIS11-0081 to the project VPS, and by the Swedish Research Council for the center of excellence CADICS. Thanks also to Jonas Unger for giving me nice, descriptive images of the VPS approach to image rendering.

**REFERENCES**


---

**Fig. 9.** 95% confidence interval for the position and angular displacements obtained by parametric CRLB. The result has been evaluated for all benchmark tracks and for different number of landmarks (3 on first row and 14 on second row) and observation noise. The distance to the wall was fixed to 7 meters.