Sensor Fusion for Augmented Reality

J. D. Hol, T. B. Schön, F. Gustafsson
Division of Automatic Control
Department of Electrical Engineering
Linköping University
SE-581 83, Linköping, Sweden
{hol,schon,fredrik}@isy.liu.se

P. J. Slycke
Xsens Technologies B.V.
Postbus 545, 7500 AM Enschede
The Netherlands
per@xsens.com

Abstract - ***In Augmented Reality (AR), the position and orientation of the camera have to be estimated with high accuracy and low latency. This nonlinear estimation problem is studied in the present paper. The proposed solution makes use of measurements from inertial sensors and computer vision. These measurements are fused using a Kalman filtering framework, incorporating a rather detailed model for the dynamics of the camera. Experiments show that the resulting filter provides good estimates of the camera motion, even during fast movements.***

Keywords: Sensor fusion, Kalman Filter, Augmented Reality, Computer Vision, Inertial Navigation.

1 Introduction

For many applications it is useful to enhance human vision with real-time computer generated virtual objects [1]. These virtual objects can for instance be used to display information aiding the user to perform real-world tasks. Typical applications range from TV and film production, to industrial maintenance, defence, medicine, education, entertainment and games. An example is shown in Figure 1, where a virtual car has been rendered into the scene.

Figure 1: An example of how AR can be used in TV production: a virtual car has been rendered into the scene.

The idea of adding virtual objects to an authentic three dimensional scene, either by displaying them in a see-through head mounted display or by superimposing them on camera images is called augmented reality [1]. For a realistic effect, the virtual objects have to be correctly aligned to the real scene. Hence, one of the key enabling technologies for AR is to be able to determine the position and orientation (pose) of the camera with high accuracy and low latency.

Prior work in this research area has mainly considered the problem in an environment which has been prepared in advance with various artificial markers, see, e.g., [2–5]. The current trend is to shift from prepared to unprepared environments, which makes the problem much harder. On the other hand, the time-consuming and hence costly procedure of preparing the environment with markers will no longer be required. Furthermore, these prepared environments seriously limit the application of AR [6]. For example, in outdoor situations it is generally not even possible to prepare the environment with markers. This problem of estimating the camera's position and orientation in an unprepared environment has previously been discussed in the literature, see, e.g., [7–11]. Furthermore, the work by [12, 13] is interesting in this context. Despite all the current research within the area, the objective of estimating the position and orientation of a camera in an unprepared environment still presents a challenging problem.

Tracking in unprepared environments requires unobtrusive sensors, i.e., the sensors have to satisfy mobility constraints and cannot modify the environment. The currently available sensor types (inertial, acoustic, magnetic, optical, radio, GPS) all have their shortcomings on for instance accuracy, robustness, stability and operating speed [14]. Hence, multiple sensors have to be combined for robust and accurate tracking.

This paper discusses an AR framework using the combination of unobtrusive inertial sensors, with a vision sensor, i.e., a camera detecting distinct features in the scene (so-called natural landmarks). Inertial sensors provide position and orientation by integrating measured accelerations and angular velocities. These estimates are very accurate on short timescales, but drift away on a longer time scale. This drift can be compensated for using computer vision, which, in itself, is not robust during fast motion. Since the inertial sensors provide accurate pose predictions, computational load required for the vision processing can be reduced by e.g., decreasing search windows or pro-

cessing at lower frame rates. This will result in minor performance degradation, but is very suitable for mobile AR applications.

A schematic illustration of the approach is given in Figure 2. The information from both sources is

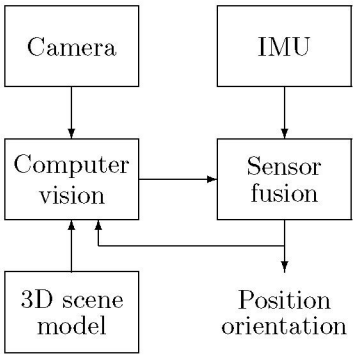

Figure 2: Schematic illustration of the approach.

fused using an Extended Kalman Filter (EKF). This method hevily relies on accurate modelling (in the form of process and observation models) of the system. The derivation and use of these models forms the main contribution of this paper.

2 Sensors

The position and orientation are determined by fusing information from a camera and an Inertial Measurement Unit (IMU). Both sensors have been integrated in a single package, shown in Figure 3. The details of

Figure 3: A hardware prototype of the MATRIS project, integrating a camera and an IMU in a single housing. It provides a hardware synchronised stream of video and inertial data.

both the IMU and the vision part will be discussed in the following sections.

2.1 Inertial Measurement Unit

The IMU is based on solid state miniature Micro-Electro-Mechanical Systems (MEMS) inertial sensors. This type of inertial sensors are primarily used in automotive applications and consumer goods. Compared to higher end MEMS inertial sensors or optical gyros, the measurements are relatively noisy and unstable and can only be used a few seconds to dead-reckon position and orientation.

The IMU is set to provide 100 Hz calibrated and temperature compensated 3D acceleration and 3D angular velocity measurements. 3D earth magnetic field data is also available, but not used. Furthermore, the IMU provides a trigger signal to the camera, which allows for exact hardware synchronisation between the sampling instances of the IMU and the camera.

The IMU sensors are individually calibrated by the manufacturer [15] to compensate for effects such as gain factor, offsets, temperature dependence, non-orthogonality, cross sensitivity, etc. However, with this type of miniature, low-cost sensor, relatively large residual sensor errors remain. Estimating these accelerometer and gyro offset errors increases the stability of the tracking. Since the inclusion of offset estimation in the models is relatively straightforward, they are suppressed for notational convenience.

2.2 Vision

The computer vision part of the AR application is based on a Kanade-Lucas-Thomasi (KLT) feature tracker and a model of the scene. The 3D scene model consists of natural features (see Figure 4). Both pixel

Figure 4: An example of a scene model and the scene it is based on.

data and 3D positions are stored for every feature. While tracking, templates are generated by warping the patches in the model according to homographies calculated from the latest prediction of the camera pose. These templates are then matched with the current camera image using the KLT tracker, similar to [13]. The vision measurements now consist of a list of 2D/3D correspondences, i.e., 3D coordinates of a feature together with its corresponding coordinates in the camera image. These correspondences can be used to estimate the camera pose.

By itself, this setup is very sensitive to even moderate motion since the search templates need to be close to reality for reliable and accurate matching. However, because of the relatively low sampling rates of the computer vision the predicted poses can be quite poor, resulting in low quality search templates. The

IMU can be used to estimate the pose quite accurately on a short time scale and hence its use drastically improves the robustness of the system.

Currently, the scene model is generated off-line using images of the scene or existing CAD models [16]. In the future Simultaneous Localisation and Mapping (SLAM) [13] will be incorporated as well.

3 Models

Several coordinate systems (shown in Figure 5) are used in order to model the setup:

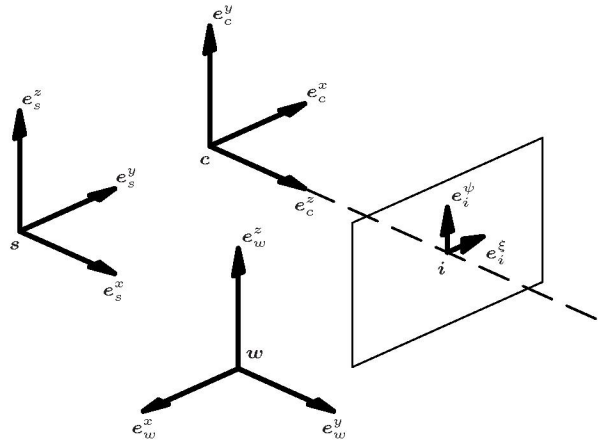

Figure 5: The world, camera, image and sensor coordinate systems with their corresponding unit vectors.

- **World (w)**: This is the reference system, fixed to earth. The (static) features of the scene are modelled in this coordinate system. Ignoring the earth's rotation, this system is an inertial frame. This coordinate system can be aligned in any way as long as the gravity vector and (if the magnetometers are used) the magnetic field vector are known.
- **Camera (c)**: The coordinate system attached to the (moving) camera. Its origin is located in the optical centre of the camera, with the z-axis along the optical axis. The camera, a projective device, takes its images in the image coordinate system. Furthermore, it has an inertial sensor attached.
- **Image (i)**: The 2D coordinate system in which the camera projects the scene. The image plane is perpendicular to the optical axis and is located at an offset (focal length) from the optical centre of the camera.
- **Sensor (s)**: This is the coordinate system of the IMU. Even though the camera and IMU are contained in a single small unit, the sensor coordinate system does not coincide with the camera coordinate system. However, the sensor is, rigidly attached to the camera with a constant translation and orientation. These parameters are determined in a calibration procedure.

The coordinate system in which a quantity is resolved in will be denoted with a superscript.

3.1 Process model

The camera pose consists of position and orientation. The position can be expressed rather straightforwardly in Cartesian coordinates. However, finding a good description for orientation is a more intricate problem and several solutions exist [17]. Unit quaternions provides an appealing solution in terms of non-singular parameters and simple dynamics. Using unit quaternions, a rotation is performed according to

$$x^a \equiv q^{ab} \odot x^b \odot \bar{q}^{ab} = q^{ab} \odot x^b \odot q^{ba}, \tag{1}$$

where $x^a, x^b, \in \mathcal{Q}_0 = \{q \in \mathbb{R}^4 : q_0 = 0\}$, $q^{ab} \in \mathcal{Q}_1 = \{q \in \mathbb{R}^4 : q \odot \bar{q} = 1\}$ and $\odot$ denotes quaternion multiplication. The notation q^{ab} is used for the rotation from the b to the a coordinate system.

The camera pose consists of the position of the camera $\boldsymbol{c}^w$ and its orientation q^{cw}. The kinematics of the camera pose are described by a set of continuous-time differential equations, briefly derived below. For a more thorough discussion of these equations, see [18].

The position of the camera $\boldsymbol{c}^w$ can be written as a vector sum (see Figure 5)

$$\boldsymbol{c}^w = \boldsymbol{s}^w + [\boldsymbol{c} - \boldsymbol{s}]^w = \boldsymbol{s}^w + q^{ws} \odot \boldsymbol{c}^s \odot q^{sw}. \tag{2}$$

Differentiating (2) with respect to time results in

$$\begin{aligned} \dot{\boldsymbol{c}}^w &= \dot{\boldsymbol{s}}^w + \boldsymbol{\omega}^w \times [\boldsymbol{c} - \boldsymbol{s}]^w \\ &= \dot{\boldsymbol{s}}^w + q^{ws} \odot [\boldsymbol{c}^s \times \boldsymbol{\omega}^s] \odot q^{sw}, \end{aligned} \tag{3}$$

where the term with $\dot{\boldsymbol{c}}^s$ has been ignored due to the fact that the sensor is rigidly attached to the camera. The accelerometer measurement can be written as

$$\boldsymbol{a}^s = q^{sw} \odot [\ddot{\boldsymbol{s}}^w - \boldsymbol{g}^w] \odot q^{ws}, \tag{4}$$

where $\boldsymbol{g}^w$ denotes the gravity vector. Rewriting (4) results in

$$\ddot{\boldsymbol{s}}^w = q^{ws} \odot \boldsymbol{a}^s \odot q^{sw} + \boldsymbol{g}^w. \tag{5}$$

Furthermore, quaternion kinematics give the following equation for the time derivative of q^{sw}

$$\dot{q}^{sw} = -\tfrac{1}{2}\boldsymbol{\omega}^s \odot q^{sw} \tag{6}$$

The derivations above can now be summarised in the following continuous-time state-space model

$$\frac{\partial}{\partial t}\begin{bmatrix} \boldsymbol{c}^w \\ \dot{\boldsymbol{s}}^w \\ q^{sw} \end{bmatrix} = \begin{bmatrix} \dot{\boldsymbol{s}}^w + q^{ws} \odot [\boldsymbol{c}^s \times \boldsymbol{\omega}^s] \odot q^{sw} \\ q^{ws} \odot \boldsymbol{a}^s \odot q^{sw} + \boldsymbol{g}^w \\ -\tfrac{1}{2}\boldsymbol{\omega}^s \odot q^{sw} \end{bmatrix}, \tag{7a}$$

which, in combination with

$$q^{cw} = q^{cs} \odot q^{sw}, \tag{7b}$$

provides a complete description of the camera pose. The non-standard state vector, $x = [\boldsymbol{c}^w, \dot{\boldsymbol{s}}^w, q^{sw}]$, as opposed to $[\boldsymbol{c}^w, \dot{\boldsymbol{c}}^w, q^{cw}]$, has the advantage that the

inertial quantities, $\boldsymbol{a}^s$ and $\boldsymbol{\omega}^s$ are measured directly by the IMU.

The discrete-time process model is now derived by integrating (7), while treating the inertial measurements as piecewise constant input signals. This dead-reckoning approach results in the following discrete-time state-space description

$$\begin{aligned}\boldsymbol{c}_{t+T}^w &= \boldsymbol{c}_t^w + T\dot{\boldsymbol{s}}_t^w + \tfrac{T^2}{2}\boldsymbol{g}^w \\ &\quad + R_t^{ws}R_t^1\boldsymbol{a}_t^s + R_t^{ws}R_t^2C^s\boldsymbol{\omega}_t^s, \qquad (8a)\\ \dot{\boldsymbol{s}}_{t+T}^w &= \boldsymbol{s}_t^w + T\boldsymbol{g}^w + R_t^{ws}R_t^1C^s\boldsymbol{\omega}_t^s, \qquad (8b)\\ q_{t+T}^{sw} &= w_t \odot q_t^{sw}, \qquad (8c)\end{aligned}$$

where T the sample time, $R^{ws}(q^{sw})$ is the rotation matrix from s to w defined as

$$R^{ws}(q^{ws}) = \begin{bmatrix} 2q_0^2+2q_1^2-1 & 2q_1q_2-2q_0q_3 & 2q_1q_3+2q_0q_2 \\ 2q_1q_2+2q_0q_3 & 2q_0^2+2q_2^2-1 & 2q_2q_3-2q_0q_1 \\ 2q_1q_3-2q_0q_2 & 2q_2q_3+2q_0q_1 & 2q_0^2+2q_3^2-1 \end{bmatrix}. \qquad (9)$$

Furthermore,

$$\begin{aligned}R^1(\boldsymbol{\omega}^s) &= TI + \tfrac{T^2}{2}\Omega^s, \qquad (10a)\\ R^2(\boldsymbol{\omega}^s) &= \tfrac{T^2}{2}I + \tfrac{T^3}{6}\Omega^s, \qquad (10b)\\ w(\boldsymbol{\omega}^s) &= \begin{bmatrix} 1 \\ -\frac{T}{2}\boldsymbol{\omega}^s \end{bmatrix}. \qquad (10c)\end{aligned}$$

Finally, I is the identity matrix, and C^s and Ω^s are skew-symmetric matrices defined according to

$$\boldsymbol{c}^s \times \boldsymbol{v} = \underbrace{\begin{bmatrix} 0 & -c_z^s & c_y^s \\ c_z^s & 0 & -c_x^s \\ -c_y^s & c_x^s & 0 \end{bmatrix}}_{C^s} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}. \qquad (11)$$

The process model (8) uses measured, hence noisy, inertial quantities as input signals. These noises are accounted for by the process noise using linearisation. Alternatively, the inertial quantities can be treated as measurement signals and not as input signals. In that case they have to be included in the state vector and consequently its dimension is increased from 10 to 16 states. An advantage with including the angular velocities and the accelerations in the state vector is that they can be predicted.

3.2 Observation model

The computer vision algorithm discussed in Section 2.2 returns a list of 2D/3D correspondences, consisting of 3D positions ($\boldsymbol{z}^w$) and corresponding image coordinates ($\boldsymbol{z}^i$). These quantities are related to each other through a camera model. When working with calibrated images, the simple pinhole camera model is applicable. It defines the map $\boldsymbol{z}^c \mapsto \boldsymbol{z}^i$, with $\boldsymbol{z}^c = [x, y, z]^T$ and $\boldsymbol{z}^i = [\xi, \psi]^T$ as

$$\begin{bmatrix} \xi \\ \psi \end{bmatrix} = \begin{bmatrix} fx/z \\ fy/z \end{bmatrix}, \qquad (12a)$$

or equivalently,

$$\boldsymbol{0} = \begin{bmatrix} z\xi - fx \\ z\psi - fy \end{bmatrix} = \begin{bmatrix} -fI_2 & \boldsymbol{z}^i \end{bmatrix} \boldsymbol{z}^c. \qquad (12b)$$

Here, f is the focal length of the camera. $\boldsymbol{z}^c$ can be calculated from its corresponding scene model entry ($\boldsymbol{z}^w$). Inserting this provides the following observation model

$$0 = \begin{bmatrix} -fI_2 & \boldsymbol{z}^i \end{bmatrix} R^{cs}R^{sw}[\boldsymbol{z}^w - \boldsymbol{c}^w] + \boldsymbol{v}_c, \qquad (13a)$$

with measurement noise $\boldsymbol{v}_c \sim N(0, \Sigma_c)$, where

$$\Sigma_c = \begin{bmatrix} -fI_2 \\ \boldsymbol{z}^{i,T} \\ z_z^c I_2 \end{bmatrix}^T \begin{bmatrix} R^{cw}\Sigma_w R^{wc} & 0 \\ 0 & \Sigma_i \end{bmatrix} \begin{bmatrix} -fI_2 \\ \boldsymbol{z}^{i,T} \\ z_z^c I_2 \end{bmatrix}. \qquad (13b)$$

The noise affecting the image coordinates and the position of the feature is assumed to be Gaussian, with zero-mean and covariances Σ_i and Σ_w, respectively. Currently, educated guesses are used for the values of these covariances. However, calculating feature and measurement dependent values is a topic under investigation.

4 Results

The process and observation models of the previous section have been implemented in an EKF [19]. The performance of this filter using measured inertial data combined with simulated correspondences will be presented in this section.

Several realistic camera motions have been carried out:

- **Pedestal**: The camera is mounted on a pedestal, which typically used for studio TV recordings. This results in very smooth motions with slow movements.
- **Hand held**: A camera man is walking around with the camera on his shoulder. Hence, the motion is still relatively smooth, but faster movements are introduced.
- **Rapid**: The camera is carried by a camera man who is running through the scene. This type of motion has relatively fast movements, since high accelerations and fast turns are present.

The three motion types described above differ in how violent the camera motion is. This is illustrated in Figure 6. The studio is also equipped with the FREE-D system [2], a conventional AR-tracking system requiring a heavy infrastructure (lots of marker on the ceiling). The pose estimates from this system can be used as ground truth data, which are used to evaluate the estimate produced by the filter proposed in this paper. To estimate the influence of various scene parameters, 2D/3D correspondences have been simulated by projecting an artificial scene onto the image plane whose position and orientation is given by the ground truth

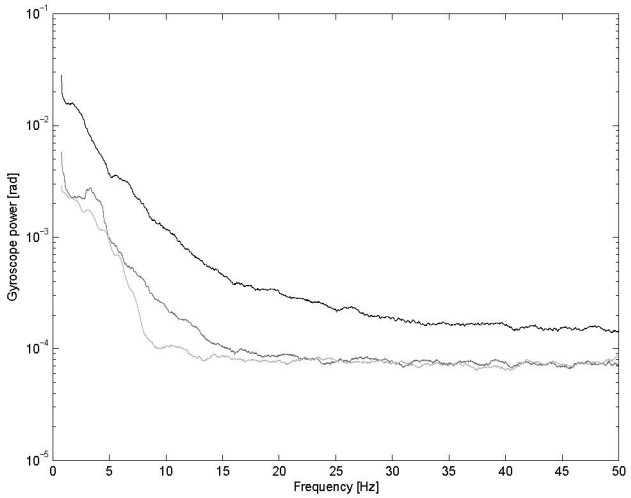

Figure 6: Power spectrum of the gyroscopes for the pedestal (light gray), hand held (dark gray) and rapid (black) motion types.

data. These virtual correspondences have been fed, with realistic noise, to the pose filter together with the captured IMU data.

The pose filter based on the proposed models performs very satisfactorily as shown in Figure 7. Note that the gaps in the error plot arise due to the fact that there is no ground truth data available during this time. This figure, generated with parameters shown in Table 1 and 2, shows the typical behaviour for the type of motion of camera man running trough the scene.

Table 1: Specifications of the sensors

IMU		
gyroscope range	$\pm$ 15.7	rad/s
gyroscope noise	0.01	rad/s
gyroscope bandwidth	40	Hz
accelerometer range	$\pm$ 20	m/s^2
accelerometer noise	0.01	m/s^2
accelerometer bw.	30	Hz
sample rate	100	Hz
Camera		
resolution	640×480	pixels
pixel size	10×7	μm/pixel
focal length	900	pixels
sample rate	50	Hz

This motion type is among the more extreme ones in terms of fast and large movements made by humans. Using only computer vision, tracking is lost almost immediately, which clearly shows the benefit of adding an IMU. For slower movements the performance is even better, as illustrated by Figure 8.

It should be noted that the described system is very sensitive to calibration parameters. For instance, small errors in the hand-eye calibration (q^{cs}) or in the intrinsic parameters of the camera will result in rapid deterioration of the tracking. Hence, design of accurate calibration methods or adaptive algorithms is of utmost importance for proper operation of the filter.

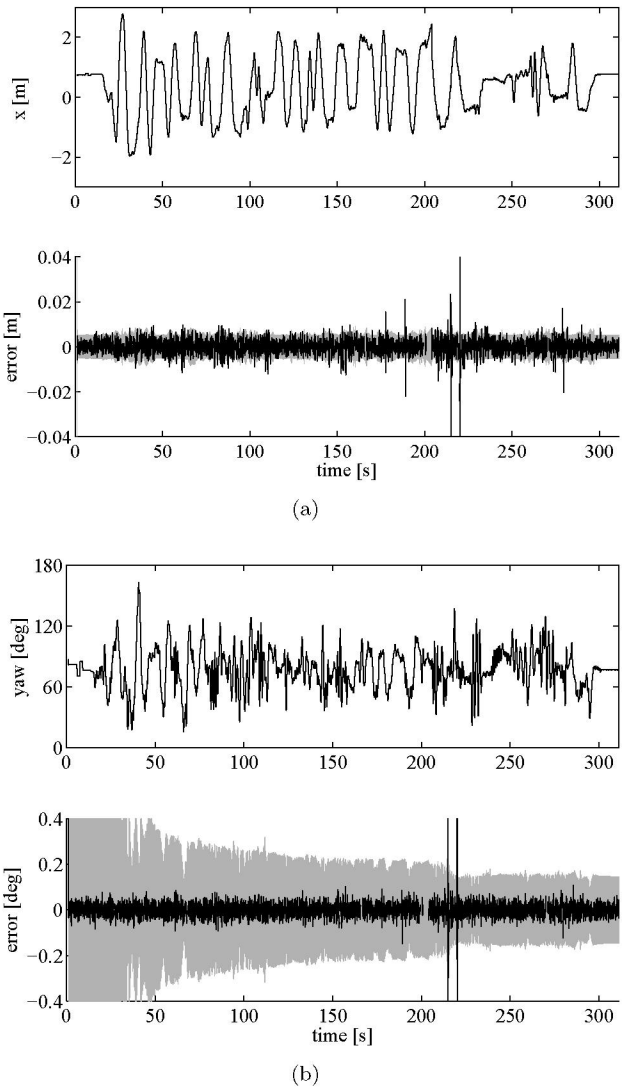

Figure 7: Pose filter estimates and errors for the rapid motion type. (a) position (x). (b) orientation (yaw). 99%-confidence levels are shown in gray. The other positions (y, z) and orientations (roll, pitch) show similar behaviour.

5 Conclusions

In this paper process and observation models are proposed for fusing computer vision and inertial measurements to obtain robust and accurate real-time camera pose tracking. The models have been implemented and tested using authentic inertial measurements and simulated 2D/3D correspondences. Comparing the results to a reference system shows stable and accurate tracking over an extended period of time for a camera that undergoes fast motion.

Even though the system works quite well, several topics require further investigation. These include design of accurate self-calibration methods, including uncertainty measures for the computer vision measurements and adding SLAM functionality for on-line scene modelling.

Table 2: Pose filter parameters

IMU sample rate	100	Hz
Camera sample rate	25	Hz
Gyroscope noise	0.014	rad/s
Gyroscope bias noise	$1{\cdot}10^{-4}$	rad/s
Accelerometer noise	0.4	m/s^2
Accelerometer bias noise	$1{\cdot}10^{-4}$	m/s^2
Scene model noise	0.01	m
Pixel noise	1	pixel
Focal length	900	pixels
Number of correspondences	30	
Feature depth	5	m

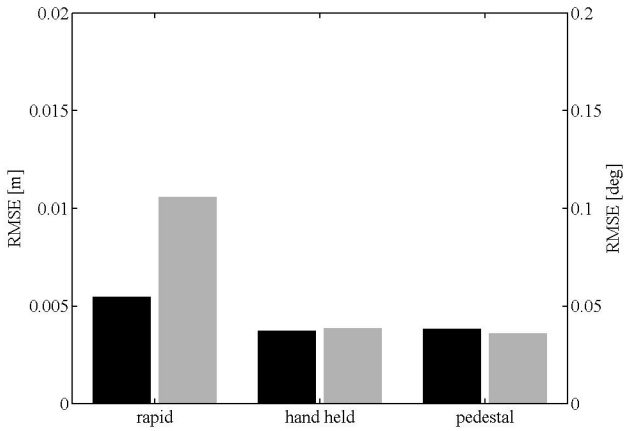

Figure 8: RMSE of position (black) and orientation (gray) for the different motion types.

Acknowledgements

This work has been performed within the MATRIS consortium, which is a sixth framework research program within the European Union (EU), contract number: IST-002013. The author would like to thank the EU for the financial support and the partners within the consortium for a fruitful collaboration this far. The MATRIS consortium aims to develop a marker-free tracking system. For more information, please visit its website, `http://www.ist-matris.org`.

References

[1] R.T. Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385, August 1997.

[2] G.A. Thomas, J. Jin, N. Niblett, and C. Urquhart. A versatile camera position measurement system for virtual reality tv production. In *International Broadcasting Conference*, pages 284–289, Amsterdam, The Netherlands, September 1997.

[3] J. Caarls, P.P. Jonker, and S. Persa. Sensor fusion for augmented reality. In *Ambient Intelligence*, volume 2875 of *Lecture Notes in Computer Science*, pages 160–176, Veldhoven, Netherlands, Nov. 2003. Springer Verlag.

[4] Y. Yokokohji, Y. Sugawara, and T. Yoshikawa. Accurate image overlay on video see-through HMDs using vision and accelerometers. In *Virtual Reality 2000 Conference*, pages 247–254, New Brunswick, NJ USA, March 2000.

[5] S. You and U. Neumann. Fusion of vision and gyro tracking for robust augmented reality registration. In *IEEE Virtual Reality 2001*, pages 71–78, Yokohama, Japan, March 2001.

[6] R. Azuma, J.W. Lee, B. Jiang, J. Park, S. You, and U. Neumann. Tracking in unprepared environments for augmented reality systems. *Computers & Graphics*, 23(6):787–793, December 1999.

[7] G. Klein and T. Drummond. Robust visual tracking for non-instrumental augmented reality. In *International Symposium on Mixed Augmented Reality*, pages 113–123, Tokyo, Japan, October 2003.

[8] S. You, U. Neumann, and R. Azuma. Hybrid inertial and vision tracking for augmented reality registration. In *IEEE Virtual Reality 1999*, pages 260–267, March 1999.

[9] G. Simon and M. Berger. Reconstructing while registering: a novel approach for markerless augmented reality. In *Proceedings of the International Symposium on Mixed and Augmented Reality*, pages 285–293, Darmstadt, Germany, September 2002.

[10] V. Lepetit, L. Vacchetti, D. Thalmann, and P. Fua. Fully automated and stable registration for augmented reality applications. In *Proceedings of the second IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 93–102, Tokyo, Japan, October 2003.

[11] Y. Genc, S. Riedel, F. Souvannavong, C. Akinlar, and N. Navab. Marker-less tracking for AR: a learning-based approach. In *Proceedings of the International Symposium on Mixed and Augmented Reality*, pages 295–304, Darmstadt, Germany, September 2002.

[12] A. J. Davison, Y. G. Cid, and N. Kita. Real-time 3D SLAM with wide-angle vision. In *Proceedings of the 5th IFAC/EUCON Symposium on Intelligent Autonomus Vehicles*, Lisaboa, Portugal, July 2004.

[13] A.J. Davison. Real-time simultanious localisation and mapping with a single camera. In *Proc. International Conference on Computer Vision*, pages 1403–1410, Nice, France, October 2003.

[14] G. Welch and E. Foxlin. Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Computer Graphics and Applications*, 22(6):24–38, Nov.-Dec. 2002.

[15] Xsens. MTi miniature attitude and heading reference system, 2006. URL `http://www.xsens.com`.

[16] R. Koch, J.-F. Evers-Senne, J.-M., Frahm, and K. Koeser. 3D reconstruction and rendering from image sequences. In *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005)*, Montreux, Switzerland, April 2005.

[17] M.D. Shuster. A survey of attitude representations. *The Journal of the Astronautical Sciences*, 41(4):439–517, October 1993.

[18] J.D. Hol. Sensor fusion for camera pose estimation. Master's thesis, University of Twente, 2005.

[19] T. Kailath, A.H. Sayed, and B. Hassibi. *Linear Estimation*. Information and System Sciences Series. Prentice Hall, Upper Saddle River, New Jersey, 2000.