International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 3 (2011) pp. 846–861

A Framework for Real-Time Scene Modelling based on Visual-Inertial Cues

Dominik Aufderheide[1,2] **and Werner Krybus**[1]

[1]South Westphalia University of Applied Sciences, Division Soest,
Institute for Computer Science, Vision and Computational Intelligence (CV&CI)
Luebecker Ring 2, 59494 Soest, Germany
{*aufderheide, krybus*}*@fh-swf.de*

[2]The University of Bolton,
School of the Built Environment and Engineering
Deane Road, Bolton BL3 5AB, U.K.
dma1bee@bolton.ac.uk

Abstract: **The self-acting generation of three-dimensional models by analysing monocular image streams from standard cameras is one fundamental problem in the field of computer vision. A prerequisite for the scene modelling is the former computation of camera poses for the different frames of the sequence. Several techniques and methodologies are introduced during recent decades to solve this classical Structure from Motion (SfM) problem, which incorporates camera egomotion estimation and subsequent recovery of scene structure. Nevertheless the applicability of those systems in real world devices and applications is still limited due to non-satisfactorily properties in terms of computational costs, accuracy and robustness. This paper suggests a novel framework for visual-inertial scene reconstruction (VISrec!) based on ideas from multi-sensor data fusion (MSDF). The integration of additional modalities (here: inertial measurements) is useful to compensate typical problems of systems which rely only on visual information.**

Keywords: Structure from Motion (SfM), Inertial sensing, Dynamic World Modeling (DWM), Multi-Sensor Data Fusion (MSDF), Kalman Filter

I. Introduction

The automatic generation of three-dimensional models has been one of the fundamental problems in computer vision for decades. Even if methods for image-based modelling are already introduced the usage of active 3D scanners is still the dominant technology in this field. Thus it is highly desirable to use monoscopic image streams which can be captured by standard digital camera devices as a base for non-invasive scene modelling. Those cameras are available at low costs and easy to handle even for a single non-professional user.
The simultaneous estimation of the camera motion and scene structure is widely known as the Structure from Motion (SfM) problem where numerous solutions were proposed and implemented. Even if the potentials of those methods are worthy for many different application fields, as Augmented Reality (AR), robot navigation or Unmanned Vehicles (UV), the applicability in real-world applications is limited due to some unsolved issues.
One of these drawbacks is the missing ability to run those classical SfM-systems in real-time due to high computational costs (see [41]) or necessary batch-type computations[1] as for classical factorisation methods (see [45]). On the other hand most SfM-methods are suffering from missing robustness of the feature detection and tracking procedures which are generating necessary input data for the recovery of shape and motion. In [6] numerous problems (namely: occlusions, depth discontinuities, low texture, repetitive patterns, etc.) of image registration are listed and analysed in the context of stereo matching. All of these problems are also considerable for SfM and many algorithms suffer from non-robust feature registration between subsequent frames of a monocular image stream. Also [43] stated that even the tracking of a subset of features is unstable in nature. By this two different problems have to be solved by SfM methods: on the one hand inaccurate localisation of matches and on the other hand a not neglectable number of complete wrong matches (outliers).
Many algorithms are also restricted to constrained type of camera movements or only a subset of possible scene types. In this context especially the necessity for a reinitialisation of the whole systems if the feature track is lost once within the sequence is a major drawback for manually operated camera systems.
In the field of mobile robotics novel methodologies and concepts for simultaneous localisation and mapping (SLAM) are recently crossing the border to real-time processing. So [21] presented the MonoSLAM-approach which is based on a single camera mounted on a moving robot. Similar ideas were used in the parallel tracking and mapping approach (PTAM) suggested by [33].
Due to these problems with classical vision-based SfM methodologies and inspired by recent developments in mobile robotics in general and SLAM in particular the general

[1]Batch-type methods are composed in such a way that the whole image sequence has to be available for the estimation of camera egomotion and scene structure. In those systems structure and motion are recovered simultaneously by solving a large-scale optimisation problem.

concept of aided Structure from Motion (aSFM) was developed. Due to the fact that the estimation of camera egomotion is a major step for 3D scene recovery the state-of-the-art for Integrated Navigation Systems (INS) is a major influence for the suggested approach of a visual-inertial approach for aSfM.
The integration of visual and inertial information was recently proposed as a methodology for full six degree of freedom (6DoF) tracking of an object's ego-motion (position and orientation) for AR applications (see [29]), navigation of UVs or mobile robot navigation [32]. Research in the field of SfM by combining visual and inertial cues is an open topic, since recently published work lacks the ability for real-time operation [19], modelling of unconstrained, dense scene reconstruction or rapid sensor movements. This paper describes a multi-modal approach for aSfM incorporating visual measurements from a single standard camera and inertial measurements from gyroscopes, accelerometers and magnetometers. The main focus lies on a naive implementation based on a two-track architecture which consists of several fusion nodes for MSDF. This paper is an extended version of the work presented in [4] with the same title.
The remainder of this paper is organised as follows: Section II gives a general introduction into the motivation for using visual-inertial cues for aSfM. Section III covers the conceptual design of the systems general architecture. In this context the idea of a two-track design is introduced and described in detail. Subsequent sections IV and V are covering the implementation of the processing of visual, respectively inertial measurements, while Section VI gives an idea about the interfaces between the visual and inertial route of the system. Finally section VII concludes the work and gives an overview about intended working packages for future research.

II. Visual-Inertial Scene Reconstruction (VISrec!) - A Motivation

Any kind of sensor measurements are uncertain and the physical property which should be determined can only be estimated with a limited level of confidence. Especially for optical measurement systems there are many possible sources of errors beside the typical random noise. Some of the typical problems of relying only on images for estimation the motion of a camera during the acquisition of a sequence were already mentioned in Section I.
The general concept of MSDF was successfully applied for example in the field of mobile robotics in recent years (see e.g. [37, 50]). One reason for the attention MSDF has covered in a wide branch of applications and scientific disciplines is the fact that a sound mathematical and formal background was developed since the mid 1990s. In this context the works of [39, 12, 24, 34, 13] are examples for particular overviews and surveys.
The major objective of applying MSDF in the field of SfM and 3D modelling is the compensation or at least attenuation of the described drawbacks of classical SfM-methods. So the incorporation of the inertial-modalities should improve the overall system performance[2] in terms of:

- **Temporal coverage** - Typical frame rates of a image processing system lie between 5 to 50 frames per second. So an update of the cameras egomotion is only available every 20 to 200 ms.
- **Accuracy** - Due to the fact that the recovered scene structure is determined based on a previously estimated relative camera position based on feature correspondences in successive frames of a sequence which are incorporated by noise and other uncertainties (see [2]) the accuracy of typical SfM-methods is limited.
- **Certainty** - Typically the certainty of SfM-algorithms is mainly influenced by the quality of the used homologous image features. For this especially the handling of outliers is an important aspect, because all the desired information is directly related to the quality of the used matches.
- **Computational costs** - All sates of the system (motion, observed scene structure) are not directly measured by a vision system, but have to be recovered from image data and adequate algorithms. As mentioned before the corresponding computational complexity leads often to the lack of ability of real time operation.

Besides these specific objectives of the MSDF-approach there are also general targets which are indirectly derived from the disadvantages of currently available 3D scanning devices as high costs, missing mobility and time consuming measurements. Thus the final system should mainly integrate standard low-cost components in a mobile easy-to-operate device.
As suggested by [39] the implementation of a MSDF-system which relies on fusion across sensors (see [59]) starts with a conceptual design based on former identification of adequate additional modalities and information channels. For this we follow the classification of relational sensor properties as given in [22, 10]. The following Table 1 gives an overview about the sensor-sensor relationships between visual and inertial measurements and clarifies the adequateness of inertial cues towards the realisation of a aSfM-system which is able to fulfil the objectives defined above.[3]
As it is indicated in Table 1 there is a asynchronous property of the different sensors observable which should lead to increased temporal coverage of the overall system. This reduces the danger of wrong or inaccurate feature matching because the stability of feature tracking is influenced in a positive manner by the higher update rate of a possible motion prediction step, which is especially important for constant velocity (CV) or constant acceleration (CA) motion models (see [53] for a definition and description of motion models in feature tracking). In this context the robustness of the feature tracking can be increased. The heterogeneous characteristics leads to a higher coverage of possible motion patterns of the camera. Furthermore the redundancy of the involved signals provides the possibility to achieve a higher accuracy of the

[2] The different categories are based on the definition of a generic notion of the qualified gain of a data fusion process in [10].

[3] The table is taken from a former publication of the author given in [5].

Table 1: Relational properties of visual and inertial sensing

Visual sensing	Inertial sensing	Property
Sensing spatial derivative with order 0 (position)	Sensing spatial derivatives with order 1 (gyroscopes - angular velocities) and order 2 (Accelerometers - translational accelerations)	Complementary
Long-term estimation for slow and smooth motion	Short-term estimation for rapid and unpredicted movements	Heterogeneous
Operating frequency: 5-30 Hz	Operating frequency: 50-1000 Hz	Asynchronous
Pose estimation from corresponding image features between successive frames	Gyroscopes: Attitude estimation from integrated rotational velocities Accelerometers: Attitude estimation (roll and pitch) from gravitational field Magnetometers: Attitude estimation from sensing earth's magnetic field	Redundant

motion estimate and as a consequence from this also accuracy of the reconstruction of the scene can be increased. Thus the integration of inertial measurements into a visual system is an adequate way for compensate typical drawbacks of the optical SfM.

Besides these specific characteristics the recent developments in the field of Micro-machined Electro-Mechanical Systems (MEMS) make it possible to integrate inertial sensors such as accelerometers, gyroscopes or magnetometers into a handheld device at low-costs. At this point it should be mentioned that the usage of MEMS sensory units is incorporating a bunch of problems based on immense drifting errors and variable biases. Those problems will be covered in Section IV in detail.

The integration of visual and inertial information was recently proposed as a methodology for full six degrees of freedom (6 DoF) tracking of an objects pose (including position and orientation) for Augmented Reality (AR) applications (see [28]), navigation of unmanned vehicles (UV) or mobile robot navigation ([32]). Research in the field of SfM by combining visual and inertial cues was recently done by [35], [18] and [17], but none of these systems can be regarded as a complete solution for on-the-fly 3D scene modelling in real-time.

The following section of this work describes the conceptual design of the proposed VISrec-system based on actual definitions from MSDF.

III. VISrec!-architecture

The first milestone in the development of a complete implementation of a VISrec!-system is the conceptual design and realisation of a dual track architecture based on ideas presented in [18, 35]. Here two separate tracks (visual and inertial routes) are considered as almost independent fusion nodes. This strategy allows the generation of two different subsystems integrating only one specific form of data sources (inertial or visual measurements). By this it is possible to compare the performances of the separate stages in a first step independently from each other and in a second step the establishment of different interfaces between both subsystems or the addition of another fusion node.

The following subsection describes the dual track architecture as a parallel fusion network by following the definitions from [39], before in the last subsection the used hardware-prototype is described in detail.

A. *Parallel fusion network*

The dual-track architecture we suggests is mainly influenced by the works of [35, 18], but we choose a formulation more closely related to the scheme of MSDF. Here each track is considered as a fusion cell as suggested by [39]. So the easiest representation of the dual-track system would be the structure shown in Figure 1. It can be seen that each fusion cell (FC) consists of at least one input which collects all sensory data. Here the inertial fusion cell (IFC) collects data from a 9-DoF inertial measuring unit which consists of a 3-DoF accelerometer unit, a 3-DoF gyroscope unit and a 3-DoF magnetometer. The visual fusion cell (VFC) collects the frames from a single camera. FCs also have additional inputs for auxiliary information (AI), which can be derived from other sources in the network, and external knowledge (EK). External knowledge collects all those additional data sources which are *a-priori* known and help to derive a higher level of abstraction which should be delivered at the output of each FC.

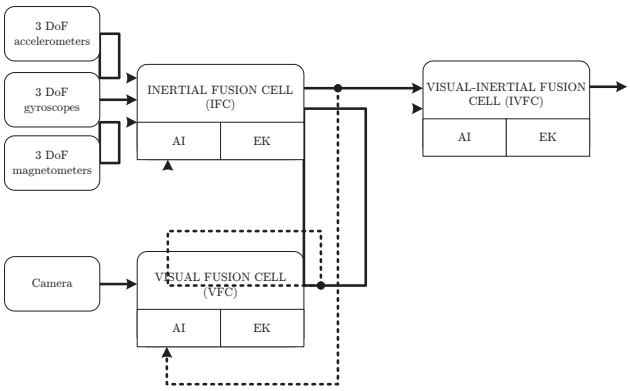


Figure. 1: Dual track system design in a representation as a parallel arrangement of fusion cells

The suggested structure contains of course different interfaces between VFC and IFC which are indicated in Figure 1 by the connections between the outputs of the two FCs and the AI-inputs of the opposite FC.

The output of the IFC will contain information about the cameras movements in a higher granularity as the raw input values (e.g. camera pose). The VFC will deliver scene structure estimates. Both signals can be collected in an additional visual-inertial fusion cell (IVFC) which realises a final refinement of structure and motion. Noteworthy there is of course the possibility that the separate FCs contain also sub-FCs for the realisation of their function.

Details about the implementation of the IFC and the VFC can be found in Sections IV and V respectively, while ideas for the realisation of the IVFC are collected in Section VI.

B. Hardware prototype

The current hardware platform used for the implementation of the shown architecture consists on a visual-inertial sensory unit build from a greyscale Unibrain Fire-i digital camera and a 9-DoF inertial unit, as shown in Figure 2.

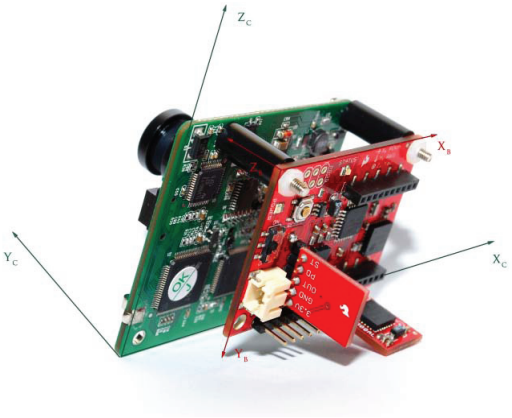


Figure. 2: Hardware prototype of the visual-inertial sensory unit

The inertial unit is inspired by the standard configuration of a multi-sensor orientation system (MODS) as defined in [49]. The used system consists of a LY530AL single-axis gyro and a LPR530AL dual-axis gyro both from STMicroelectronics, which are measuring the rotational velocities about the three main axis of the inertial coordinate system ICS (see Figure 2). The accelerations for translational movements are measured by a triple-axis accelerometer ADXL345 from Analog Devices. Finally a 3-DoF magnetometer from Honeywell (HMC5843) is used to measure the earth gravitational field. All IMU sensors are connected to a micro controller (ATMega 328) which is responsible for initialisation, signal conditioning and communication. The data from all sensors are transferred from the MODS to a standard PC. The digital camera is connected to a PC by using a standard Firewire-interface (IEEE1394).
The whole implementation of the different FCs is realised on the standard PC, as described in the subsequent sections.

IV. Inertial Fusion Cell (IFC)

The inertial route contains all the steps which are necessary to determine position and orientation of the MODS (which is rigidly attached to the camera). As already indicated in Section III-B the used MODS consists of three orthogonal arranged accelerometers measuring a three dimensional acceleration $\mathbf{a}^b = [a_x\, a_y\, a_z]^T$ normalised with the gravitational acceleration constant g. Here b indicates the actual body coordinate system in which the entities are measured. In addition three gyroscopes measuring the corresponding angular velocities $\omega^b = [\omega_x\, \omega_y\, \omega_z]^T$ around the sensitivity axes of the accelerometers. Also magnetometers with three perpendicular sensitivity axes are used to sense the earth's magnetic field $\mathbf{m}^b = [m_x\, m_y\, m_z]^T$.
Classical approaches for inertial navigation are stable-platform systems which are isolated from any external rotational motion by specialised mechanical platforms. In comparison to those classical stable platform systems the MEMS

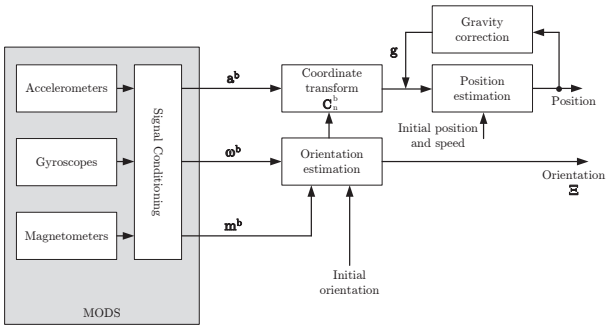


Figure. 3: Computational elements of an INS

sensors are mounted rigidly to the device (here: the camera). In such a strapdown system it is necessary to transform the measured quantities of the accelerometers into a global coordinate system by using known orientations computed from gyroscope measurements.
In general the mechanisation of a strapdown inertial navigation systems (INS) can be described by the computational elements indicated in Figure 3.
The main problem with this classical framework is that location is determined by integrating of measurements from gyros (orientation) and accelerometers (position). Due to superimposed sensor drift and noise, which is especially for MEMS devices not neglectable, the errors for the egomotion estimation tend to grow unbounded. Besides that the danger of ambiguities during initialisation of initial conditions is given. It was shown e.g. by [16] that a combination with magnetometers can help to reduce drift error.
The calibration of IMUs can be realised by moving the IMU with specialised mechanical platforms or industrial robots to known orientations with precisely controlled accelerations and rotational velocities. This provides a possibility for the determination of calibration parameters for a given sensor model and allows a signal correction.
So the final framework for pose estimation considers two steps: an orientation estimation and a position estimation as shown in Figure 4. In terms of FCs the whole procedure can again described as a sub-network of FCs which are located inside the inertial fusion cell of the overall system design, as indicated in Figure 1. In comparison to the classical strapdown mechanisation as described e.g. in [51, 60] the suggested approach here incorporates also the accelerometers for orientation estimation. The suggested fusion network is visualised in the following figure, whereat the different subfusion processes are described in subsections IV-A and IV-B.

A. Fusion for orientation

The estimation of the orientation of the MODS is realised in most approaches just based on information from the magnetometer and the gyroscopes. The most simple approach is the implementation based only on a single integrator. Due to the fact that the MEMS-implementation of the gyroscopes is suffering from an immense drifting error such a system is only stable for short-term sequences. The following Figure 5 gives an indication for the accumulated drifting error over time, while on the left hand side a comparison between the

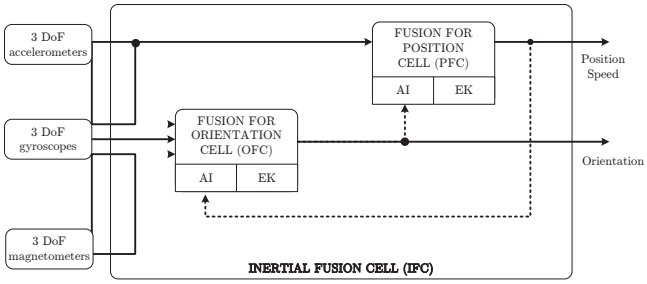


Figure. 4: System design of the inertial fusion cell (IFC)

true and the determined angle (here: roll) is shown and on the right hand side the corresponding residual. It can be easily seen that over time the error is accumulated over time.

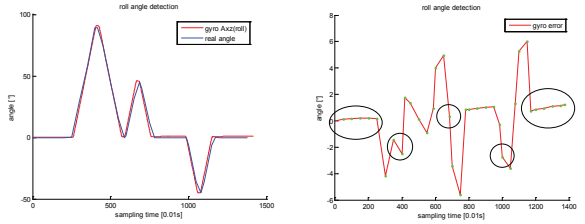


Figure. 5: Drifting error of gyroscope measurements

The general idea for compensating the drift error of the gyroscopes is based on using the accelerometer as an additional attitude sensor for generating redundant information. Due to the fact that the 3-DoF accelerometer measures not only (external) translational motion, but also the influence of the gravity it is possible to calculate the attitude based on the single components of the measured acceleration. This is of course only true if no external force is accelerating the sensor. So there are to questions which have to be answered: 1. How it is possible to calculate the attitude from accelerometer measurements? and 2. How external translational motion can be handled? Both problems can be solved by following a two-stage switching behaviour inspired by work presented in [47]. At this point it should be pointed out that measurements from the accelerometers can only provide roll and pitch angle and the heading angle has to be derived by using the magnetometer instead.
Figure 6 gives an illustration about the geometrical relations between measured accelerations due to gravity and the roll and pitch angle of the attitude. By this it follows that the angles can be determined by following relations:

$$\theta = arctan2\left(a_x^2, \sqrt{(a_y + a_z)^2}\right) \tag{1}$$

$$\phi = arctan2\left(a_y^2, \sqrt{(a_x + a_z)^2}\right) \tag{2}$$

The missing heading angle can be recovered by using the readings from the magnetometer and the already determined roll and pitch angles. Here it is important to consider that the measured elements of the earth magnetic field have to be transformed to the local horizontal plane (tilt compensation). Figure 7 is indicating the corresponding relations as shown in [16]:

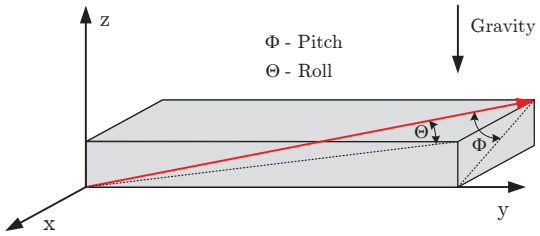


Figure. 6: Geometrical relations between measured accelerations due to gravity and the roll and pitch angle of the attitude

$$\begin{aligned} X_h &= m_x \cdot c\varphi + m_y \cdot s\theta \cdot s\varphi - m_z \cdot s\theta \cdot s\varphi \\ Y_h &= m_y \cdot c\theta + m_z \cdot s\theta \\ \psi &= \arctan 2\,(Y_h, X_h) \end{aligned} \tag{3}$$

Based on these findings a discrete Kalman filter bank (DKF-bank) is implemented which is responsible for the estimation of all three angles of Ξ. For the pitch and the roll angle the same DKF-architecture is used, as indicated in Figure 8-(a). In comparison to that the heading angle is estimated by a alternative architecture as shown in Figure 8-(b).
All DKFs are mainly based on the classical structure of a Kalman filter (see [12]) which consists of a first prediction of states and subsequent correction, where the two states are the unknown angle ξ and the bias of the gyroscope b_{gyro}. The Kalman filtering itself is composed from the following classical steps, whereat the following descriptions are simplified to a single angle ξ.

1) Computation of an a priori state estimate $\mathbf{x_{k+1}^-}$

As already mentioned the hidden states of the system are $\mathbf{x} = [\xi, \mathbf{b_{gyro}}]^\mathbf{T}$. The a priori estimates are computed by following the following relations:

$$\begin{aligned} \widehat{\omega}_{k+1} &= \omega_{k+1} - b_{gyro_k} \\ \xi_{k+1} &= \xi_k + \int \widehat{\omega}_{k+1} dt \\ b_{gyro_{k+1}} &= b_{gyro_k} \end{aligned} \tag{4}$$

Here the actual measurements from the gyroscopes ω_{k+1} are corrected by the actually estimated bias b_{gyro_k} from the former iteration, before the actual angle ξ_{k+1} is computed.

2) Computation of a priori error covariance matrix $\mathbf{P_{k+1}^-}$

The a priori covariance matrix is calculated by incorporating the Jacobi matrix $\mathbf{A}$ of the states and the process noise covariance matrix $\mathbf{Q_K}$ as follows:

$$\mathbf{P_{k+1}^-} = \mathbf{A} \cdot \mathbf{P_k} \cdot \mathbf{A^T} + \mathbf{Q_K} \tag{5}$$

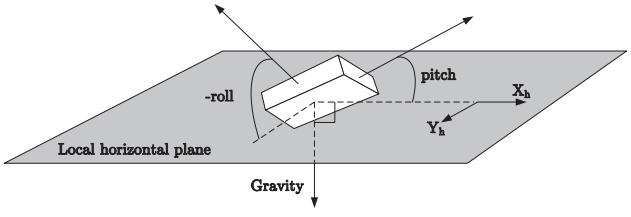


Figure. 7: Local horizontal plane as a reference

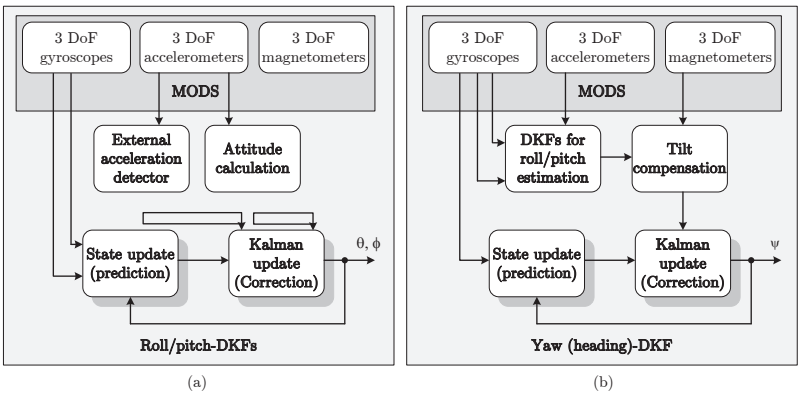


Figure. 8: (a) - Discrete Kalman filter (DKF) for estimation of roll and pitch angles based on gyroscope and accelerometer measurements; (b) - DKF for estimation of yaw (heading) angle from gyroscope and magnetometer measurements

The two steps 1) and 2) are the elements of the prediction step as indicated in Figure 8.

3) Computation of Kalman gain $\mathbf{K}_{k+1}$

As a prerequisite for computing the a posteriori state estimate the Kalman gain $\mathbf{K}_{k+1}$ has to be determined by following Equation 6.

$$K_{k+1} = \mathbf{P}^{-}_{\mathbf{k+1}} \cdot \mathbf{H}^{T}_{k+1} \cdot \left(\mathbf{H}_{k+1} \cdot \mathbf{P}^{-}_{\mathbf{k+1}} \cdot \mathbf{H}^{T}_{k+1} + \mathbf{R}_{k+1}\right)^{-1} \tag{6}$$

4) Computation of a posteriori state estimate $\mathbf{x}^{+}_{\mathbf{k+1}}$

The state estimate can now be corrected by using the calculated Kalman gain $\mathbf{K}_{k+1}$. Instead of incorporating the actual measurements as in the classical Kalman structure the suggested approach is based on the computation of an angle difference $\Delta\xi$. The difference is a comparison of the angle calculated from the gyroscope measures and the corresponding attitude as derived from the accelerometers, respectively the heading angle from the magnetometer, as already introduced in the introduction of this chapter. So the relation for $\mathbf{x}^{+}_{\mathbf{k+1}}$ can be formulated as:

$$\mathbf{x}^{+}_{\mathbf{k+1}} = \mathbf{x}^{-}_{\mathbf{k+1}} - \mathbf{K}_{k+1} \cdot \Delta\xi \tag{7}$$

At this point it is important to consider the fact that the attitude measurements from the accelerometers are only reliable if there is no external translational motion. For this an external acceleration detection mechanism is also part of the fusion procedure. For this reason the following condition (see [47]) is evaluated continuously:

$$\|\mathbf{a}\| = \sqrt{(a_x^2 + a_y^2 + a_z^2)} \stackrel{!}{=} 1 \tag{8}$$

If the relation is fulfilled there is no external acceleration and the estimation of the attitude from accelerometers is more reliable than the one computed from rotational velocities as provided by the gyroscopes. Noteworthy for real sensors an adequate threshold $\epsilon_{\mathbf{g}}$ is introduced to define an allowed variation from this ideal case. If the camera is not at rest the observation variance for the gyroscope data σ_g^2 is set to zero. So by incorporating the magnitude of the acceleration measurements as $\|\mathbf{a}\|$ and the earth gravitational field $\mathbf{g} = [0, 0, -g]^T$ the observation variance can be defined by following Equation 9.

$$\sigma_g^2 = \begin{cases} \sigma_g^2, & \|\mathbf{a}\| - \|\mathbf{g}\| < \varepsilon_{\mathbf{g}} \\ 0, & otherwise \end{cases} \tag{9}$$

A similar approach is chosen to overcome the problems with the magnetometer measurements in magnetically distorted environments for the DKF for the heading angle. Instead of gravity $\mathbf{g}$ the magnitude of the earth magnetic field $\mathbf{m}$ is evaluated as shown in the following relation[4]:

$$\sigma_g^2 = \begin{cases} \sigma_g^2, & \|\mathbf{m}\| - m_{des} < \varepsilon_{\mathbf{m}} \\ 0, & otherwise \end{cases} \tag{10}$$

5) Computation of posteriori error covariance matrix $\mathbf{P}^{+}_{\mathbf{k+1}}$

Finally the error covariance matrix is updated in the following way:

$$\mathbf{P}^{+}_{k+1} = \mathbf{P}^{-}_{k+1} - \mathbf{K}_{k+1} \cdot \mathbf{H}_{k+1} \cdot \mathbf{P}^{-}_{k+1} \tag{11}$$

B. Fusion for position

At this point the orientation of the camera is known and by following the classical strapdown mechanisation, as shown in Figure 4, the next steps for position estimation consist of the transformation from body-coordinate frame to the global navigation coordinate system and the double integration of accelerometer measurements.
In the actual configuration of the system all measurements are resolved in a body-coordinate frame, rather than a global inertial system. Hence, the position $\mathbf{p}$ can only be obtained by double integration of the body accelerations $\mathbf{a}$, when a known orientation $\mathbf{\Xi} = [\phi\, \theta\, \psi]^T$ is available that allows a rotation from body frame B to reference (or navigation) frame N by using the direct cosine matrix (DCM) $\mathbf{C}_n^b$, defined as follows[5]:

[4] m_{des} describes the magnitude of the earth's magnetic field (e.g. 48 μT in Western Europe)

[5] For simplification: $s\alpha = sin(\alpha)$ and $c\beta = cos(\beta)$

$$\mathbf{C}_n^b(\mathbf{q}) = \frac{1}{\sqrt{q_4^2 + \|\mathbf{e}\|^2}} \cdot \begin{bmatrix} q_1^2 - q_2^2 - q_3^2 + q_4^2 & 2(q_1q_2 + q_3q_4) & 2(q_1q_3 - q_2q_4) \\ 2(q_1q_2 - q_3q_4) & -q_1^2 + q_2^2 - q_3^2 + q_4^2 & 2(q_2q_3 + q_1q_4) \\ 2(q_1q_3 + q_2q_4) & 2(q_2q_3 - q_1q_4) & -q_1^2 - q_2^2 + q_3^2 + q_4^2 \end{bmatrix} \tag{12}$$

$$\mathbf{C}_n^b = \begin{bmatrix} c\theta c\psi & s\varphi s\theta c\psi - c\varphi s\psi & c\varphi s\theta c\psi + s\varphi s\psi \\ c\theta s\psi & s\varphi s\theta s\psi + c\varphi c\psi & c\varphi s\theta s\psi - s\varphi c\psi \\ -s\theta & s\varphi c\theta & c\varphi c\theta \end{bmatrix} \tag{13}$$

The DCM can also be expressed in terms of an orientation quaternion $\mathbf{q} = [\mathbf{e}^T, q_4]^T$, where $\mathbf{e} = [q_1, q_2, q_3]^T$ describes the vector part and q_4 is the scalar part of $\mathbf{q}$. Equation 12 shows the relation between $\mathbf{C}_n^b$ and a computed $\mathbf{q}$. A detailed introduction in quaternions for representing rotations can be found in [52].
The actual position is computed by double integration of accelerometer measurements.

C. Evaluation

The evaluation of the orientation estimation was realised by attaching the VISrec!-prototype to an industrial robot platform. A ABB IRB1400 industrial robot as shown in Figure 9 was used to generate different motion patterns where the ground truth data is known.

Figure. 9: ABB industrial robot for determination of ground truth motion data

At this point the following figure gives just an impression about the performance of the DKF approach in comparison to the usage of gyroscopes alone, whereat the roll angle is shown for vibration without motion (Figure 10-(a)) and for a specified rotation pattern (Figure 10-(b)). It can be clearly seen that the DKF is improving the situation enormously in terms of long-time stability and accuracy by the incorporation of accelerometer attitude measurements.
The same experiment is also done for the yaw angle and using the magnetometer as an additional information source. By observation of the residuals (Figure 11-right) it can be determined that the accuracy of orientation estimation can also be increased for this case. Noteworthy the performance for the heading angle is of course dependent of the environmental magnetic disturbances during the measurements which

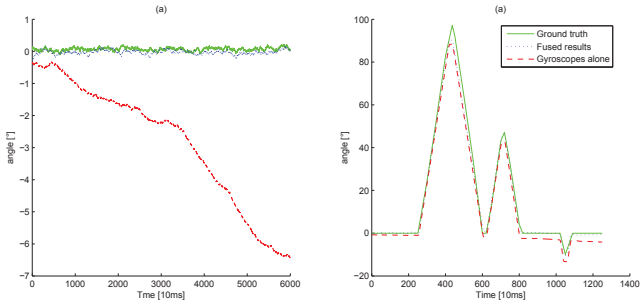


Figure. 10: Results of orientation estimation for the roll angle for (a): no rotations only noise and vibration; (b): rotation pattern

was one sever problem during the data acquisition near the moving industrial robot. Due to the definition from Equation 10 the system is relying completely on gyroscope measurements if the magnetic disturbance exceeds over a specified threshold.

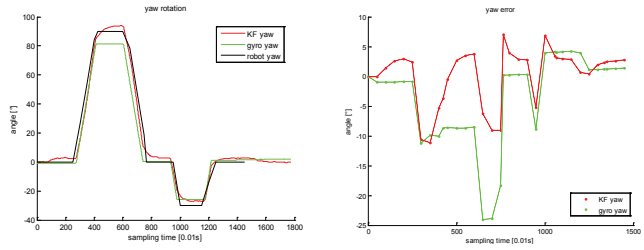


Figure. 11: Results of orientation estimation for the yaw angle for a specified motion pattern left: estimated angle; right: residual

V. Visual Fusion Cell (VFC)

For the VFC classical SfM algorithms have to be reconsidered and evaluated for their applicability in the given context, but most of those methods are fundamentally offline in nature (see e.g. [46]) due to their structure based on batch-computation. Especially those methods proposed for 3D model generation are mostly based on analysing a complete given image sequence and not successive frames. An example of such an approach can be found in [23]. Recently new approaches for SLAM, as those proposed by [21], are highly focused on the ability for high frame-rate real-time performance motivated by the intended usage in mobile robotics, but the focus is not a dense and accurate 3D reconstruction of the scene but rather a robust localisation. Thus this methodology is also labelled as visual odometry (VO). For the implementation of a mobile on-the-fly scene acquisition device the recently developed methods for SfM and SLAM have to be combined, due to the goal of a sequentially growing scene structure model which consists of reliable 3D feature points

acquired in real-time during camera motion. Figure 12 illustrates the main stages of the VFC as described in the remainder of this section.

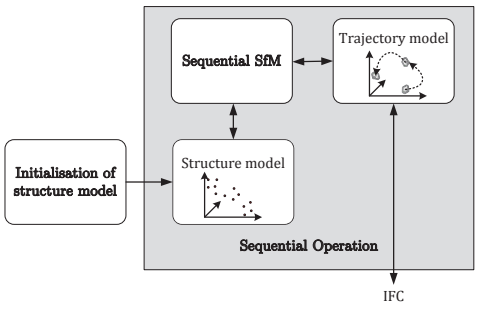


Figure. 12: Overview of the elements of the visual fusion cell

A. *Initialisation of structure model*

The VFC of the two-track system design consists of two separate steps: the initialisation of the structure model and the sequential SfM. The initial structure model is generated at the beginning of the data acquisition and can be used during the sequential SfM phase to estimate the absolute pose of the camera.

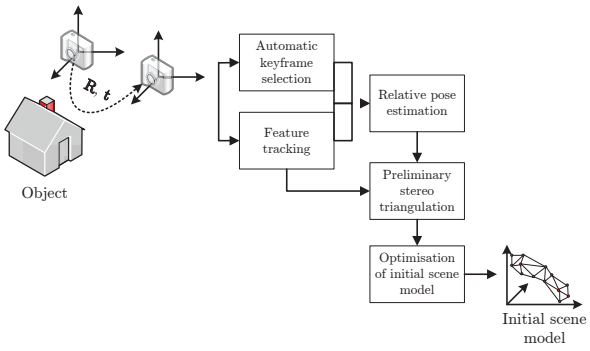


Figure. 13: Elements of the initialisation phase for the VisR

The Figure 13 gives an overview about the different elements of the initialisation phase as described in the following subsections, starting with the acquisition of the initial sequence in section V-A.1, followed by the recovery of motion information for initial keyframes in sections V-A.2 and V-A.3 and the subsequent generation of an initial model, as described in sections V-A.5 and V-A.5. Finally a bundle adjustment scheme, as described in subsection V-A.6 is used for optimisation of the initial model.

1) *Acquisition of the initial sequence*

Due to the fact that the usage of the five-point relative pose algorithm as proposed by [42] in 2004 leads to a scale ambiguity for the translational motion it is necessary for the generation of the initial structure model to capture an initial sequence where the translational motion between the first and the last frame is approximately known. This can be done in the final scheme of the two-track system design to use position information from the IFC. For the first tests, as explained here, a fixed translational motion of 600 mm is assumed and the operator of the camera has to manually start and finish the acquisition of the initial sequence, e.g. by pressing a button. By incorporating this initial guess of the translational motion it is possible to get a more adequate initial reconstruction of the feature points which is an important factor for the final bundle adjustment, because only ”good” initial values guarantee an optimal convergence of the nonlinear optimisation routine.
The initial sequence is acquired during the camera is moved in front of the object by approximately 600 mm in one direction. Figure 14 illustrates the acquisition of the initial sequence which contains of n frames. The overall translation between the first frame of the sequence $\mathbf{I_1}$ and the last one ($\mathbf{I_n}$) is assumed as $\mathbf{t_{13}} = [t_{init}, 0, 0]^T$, where t_{init} represents a fixed known translation between the first and the last frame of the initial video stream.

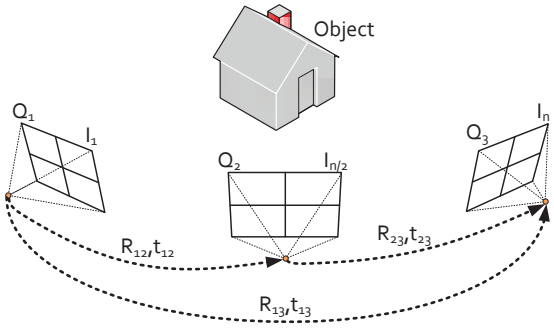


Figure. 14: Acquisition of the initial sequence

From the overall frames of the initial sequence three keyframes $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ are selected.
The three keyframes are used subsequently for the estimation of the relative pose and the partial stereo reconstruction of the observed object as described in the following sections.

2) *Relative pose estimation between key frames*

The three first keyframes of the initialisation sequence are used to generate two relative pose estimates by following general five-point relative pose algorithms as the one proposed by [42]. For this at least five points ($\mathbf{P_i}$) have to be matched successfully between two of the three keyframes.
The general problem of relative pose estimation based on a set of 2D/2D correspondences can be formulated as the recovery of time-varying parameters of a cameras egomotion $\mathbf{R}_k, \mathbf{t}_k$ from corresponding image feature coordinates $[u_{i,k}, v_{i,k}]^T$. In this context it is necessary to distinguish two different setups: the calibrated or uncalibrated camera setup. The relative pose parameters $\mathbf{R}_k, \mathbf{t}_k$ are directly related to the essential matrix $\mathbf{E}$ as defined as follows:

$$\mathbf{E_k} = \mathbf{R_k} \left[\mathbf{t_k}\right]_\times \tag{14}$$

The essential matrix describes the general epipolar relations for a stereo image pair. Here $\mathbf{X}_i$ describes a point in the world coordinate system which is imaged on the two image planes Π and Π'. So two corresponding image feature points are localised at $\mathbf{x}_i$, respectively $\mathbf{x}'_i$.
In general for an image point in homogeneous coordinates $\mathbf{x} = [u\, v\, 1]^T$ in image $\mathbf{I}$ and an corresponding image point

$\mathbf{x}' = [u'\, v'\, 1]^T$ in image $\mathbf{I}'$ the simplified epipolar constraint as shown in the following equation is true:

$$\mathbf{q}'^T \mathbf{E} \mathbf{q} = 0 \tag{15}$$

Where $\mathbf{q}$ and $\mathbf{q}'$ are computed by multiplication of the image points with the inverse of the predetermined calibration matrices $\mathbf{K}$ and $\mathbf{K}'$ of the camera. Those coordinates are called camera normalised coordinates.

$$\mathbf{q} = \mathbf{K}^{-1}\mathbf{x} \text{ and } \mathbf{q}' = \mathbf{K}'^{-1}\mathbf{x}' \tag{16}$$

The intrinsic calibration matrices $\mathbf{K}$ and $\mathbf{K}'$ are determined within a prior calibration routine following the procedure of the Camera Calibration Toolbox of Bouguet.
$\mathbf{K}$ is in general composed as shown in Equation 17, where the parameters u_o and v_0 describe a translation along the image plane and α_u, α_v and γ describe scale changes along the image axes and a rotation in the image plane (see [9]).

$$\mathbf{K} = \begin{bmatrix} \alpha_u & \gamma & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{17}$$

The definition in equation 16 shows also the relation between the essential and the fundamental matrix $\mathbf{F}$:

$$\mathbf{F} = \mathbf{K}^{-T}\mathbf{E}\mathbf{K}'^{-1} \tag{18}$$

$\mathbf{F}$ can be used to define the general epipolar constraint as shown in Equation 19.

$$\mathbf{x}'^T \mathbf{F} \mathbf{x} = 0 \tag{19}$$

One important constraint for estimation of both essential and fundamental matrix is the fact that both matrices are singular. So their determinants are both zero:

$$det(\mathbf{F}) = \mathbf{0} \text{ and } \mathbf{det}(\mathbf{E}) = \mathbf{0} \tag{20}$$

During the last decades many different algorithms are dealing with estimating both the essential and fundamental matrix from point correspondences. The approach which is mostly used in literature over years is the so called eight point algorithm which is widely used to estimate $\mathbf{F}$ and subsequently derive $\mathbf{E}$ by following Equation 18. A detailed description can be found e.g. in [26].
By using the additional constraint from Equation 20 it is possible to reduce the minimal number of points for estimating $\mathbf{E}$ to seven. As indicated by [26] it is necessary to normalise the point correspondences due to the dependency of the estimation techniques to the range of the measured values. For this case Hartley suggested an isotropic scaling which can be summarised as translate all points (in inhomogeneous coordinates) so that their mean coordinate is at the origin and scale the points that the average distance from the origin is equal to $\sqrt{(}2)$.
It was shown by [44], that an additional property, as shown in Equation 21 of the essential matrix, which can be derived from the fact that the two non-zero singular values of $\mathbf{E}$ are equal, can be used to reduce the sufficient number of points to six (see [44]), respectively five (see [42]).

$$\mathbf{E}\mathbf{E}^T\mathbf{E} - \frac{1}{2} trace\left(\mathbf{E}\mathbf{E}^T\right)\mathbf{E} = 0 \tag{21}$$

It was shown in an experimental evaluation by [48] that the usage of five-point algorithms outperforms other techniques, especially for noisy data. Even if [14] suggested a combination of an eight-point and an five-point estimator as the optimal solution for robust relative pose, the current approach considers the five-point relative pose estimator as suggested by [42]. A experimental evaluation of different techniques is provided in [3]. Each pair of corresponding points in the images $\mathbf{x}$ is leading to one equation following the constraint shown in Equation 15. [42] suggests the formulation $\tilde{\mathbf{q}}^T\tilde{\mathbf{E}} = 0$, as shown in Equation 26.
For all five point correspondences the following 5x9 data matrix $\tilde{\mathbf{Q}}$ can be obtained:

$$\tilde{\mathbf{Q}} = \begin{bmatrix} \tilde{\mathbf{q}}^1_{[1]} & \cdots & \tilde{\mathbf{q}}^1_{[9]} \\ \vdots & \vdots & \vdots \\ \tilde{\mathbf{q}}^5_{[1]} & \cdots & \tilde{\mathbf{q}}^5_{[1]} \end{bmatrix} \tag{23}$$

The solution for $\mathbf{E}$ is found by first decomposing $\tilde{\mathbf{Q}}$ by singular value decomposition (SVD) (see [14]) or QR-factorisation (see [42]) to compute the null space. The null space is leading to vectors $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}$ and $\tilde{\mathbf{D}}$. Than the following linear combination is leading to the essential matrix:

$$\mathbf{E} = a \cdot \tilde{\mathbf{A}} + b \cdot \tilde{\mathbf{B}} + c \cdot \tilde{\mathbf{C}} + d \cdot \tilde{\mathbf{D}} \tag{24}$$

It should be stated here that the four scalar values a,b,c and d are just defined up to a common scale, so it can be assumed that $d = 1$. Substituting Equation 24 into the constraints as shown in Equations 19 and 21 the problem can be formulated as ten polynomials of third degree. Nister suggested an algorithm for solving the problem to recover the unknowns of the system and recovering the essential matrix $\mathbf{E}$, whereat up to ten solutions are possible. In recent years different methods for the final estimation of $\mathbf{E}$ were suggested in literature. The original algorithm proposed by Nister in [42] uses *Sturm sequences* to solve a univariate formulation of the problem. Later [55] proposed a more efficient procedure based on *Groebner bases*. It was suggested by [61] that a formulation as a polynomial eigenvalue problem is more straightforward and leads to solutions which are numerically more stable. The different methods were evaluated in terms of accuracy and robustness against noise for the current project (see [3]).
In most cases the feature detection and matching routine will produce more than the minimum set of five correct point correspondences. In those cases the "best" solution can be found by evaluating a defined error metric. Different kinds of error metrics are defined in literature. So [48] suggests the usage of the Sampson error metric d_e over all matches ℓ, which should be minimal for the correct solution of $\mathbf{E}$ and can be defined as follows:

$$d_e = \sum_{K=1}^{\ell} \frac{\left(\tilde{\mathbf{x}}_{k'}^T \mathbf{E} \tilde{\mathbf{x}}_k\right)}{[\mathbf{E}\tilde{\mathbf{x}}_k]_x^2 + [\mathbf{E}\tilde{\mathbf{x}}_k]_y^2 + [\mathbf{E}^T\tilde{\mathbf{x}}'_k]_x^2 + [\mathbf{E}^T\tilde{\mathbf{x}}'_k]_y^2} \tag{25}$$

[26] uses the classic algebraic error based on the simplified epipolar constraint as already defined in Equation 15. Another error metric is the symmetric squared geometric error, as suggested by [14]:

$$\begin{aligned}\tilde{\mathbf{q}} &= \begin{pmatrix} \tilde{\mathbf{x}}_{[1]}\tilde{\mathbf{x}}'_{[1]} & \tilde{\mathbf{x}}_{[2]}\tilde{\mathbf{x}}'_{[1]} & \tilde{\mathbf{x}}_{[3]}\tilde{\mathbf{x}}'_{[1]} & \tilde{\mathbf{x}}_{[1]}\tilde{\mathbf{x}}'_{[2]} & \tilde{\mathbf{x}}_{[2]}\tilde{\mathbf{x}}'_{[2]} & \tilde{\mathbf{x}}_{[3]}\tilde{\mathbf{x}}'_{[2]} & \tilde{\mathbf{x}}_{[1]}\tilde{\mathbf{x}}'_{[3]} & \tilde{\mathbf{x}}_{[2]}\tilde{\mathbf{x}}'_{[3]} & \tilde{\mathbf{x}}_{[3]}\tilde{\mathbf{x}}'_{[3]} \end{pmatrix}^T \\ \tilde{\mathbf{E}} &= \begin{pmatrix} \mathbf{E}_{[1,1]} & \mathbf{E}_{[1,2]} & \mathbf{E}_{[1,3]} & \mathbf{E}_{[2,1]} & \mathbf{E}_{[2,2]} & \mathbf{E}_{[2,3]} & \mathbf{E}_{[3,1]} & \mathbf{E}_{[3,2]} & \mathbf{E}_{[3,3]} \end{pmatrix}^T \end{aligned} \tag{26}$$

$$d_{ssg} = \frac{\left(\tilde{\mathbf{x}}_{k'}^T \mathbf{E}\tilde{\mathbf{x}}_k\right)^2}{\left[\mathbf{E}\tilde{\mathbf{x}}_k\right]_x^2 + \left[\mathbf{E}\tilde{\mathbf{x}}_k\right]_y^2} + \frac{\left(\tilde{\mathbf{x}}_{k'}^T \mathbf{E}\tilde{\mathbf{x}}_k\right)^2}{\left[\mathbf{E}^T\tilde{\mathbf{x}}'_k\right]_x^2 + \left[\mathbf{E}^T\tilde{\mathbf{x}}'_k\right]_y^2} \tag{26}$$

3) Recovering motion parameters

Once the essential matrix is known the egomotion of the camera between two successive frames can be retrieved from $\mathbf{E}$. It has to be stated here that $\mathbf{E}$ can just be recovered up to scale. There is also an ambiguity, such that there are four possible solutions regarding the rotation matrix and the translation vector.
The first step in determining $\mathbf{R}$ and $\mathbf{t}$ from $\mathbf{E}$ is the computation of the singular value decomposition (SVD) of the essential matrix:

$$\mathbf{E} \sim \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{27}$$

As it was shown in [58, 26] the four possible solutions $\mathbf{R}$ and $\mathbf{t}$ can be composed based on two different solutions for the rotation matrix $\mathbf{R}_a$, $\mathbf{R}_b$ and two different solutions for the translation $\mathbf{t}_a$, $\mathbf{t}_b$ as follows: $\{\mathbf{R}_a, \mathbf{t}_a\}$, $\{\mathbf{R}_b, \mathbf{t}_b\}$, $\{\mathbf{R}_a, \mathbf{t}_b\}$ and $\{\mathbf{R}_b, \mathbf{t}_a\}$.
The definition of the solutions is based on the following definitions for $\mathbf{t}_a$ and $\mathbf{t}_b$:

$$\begin{aligned}\mathbf{t}_a &\equiv \begin{bmatrix} \mathbf{U}_{[1,3]} & \mathbf{U}_{[2,3]} & \mathbf{U}_{[3,3]} \end{bmatrix}^T \\ \mathbf{t}_b &\equiv -1 \cdot \begin{bmatrix} \mathbf{U}_{[1,3]} & \mathbf{U}_{[2,3]} & \mathbf{U}_{[3,3]} \end{bmatrix}^T \end{aligned} \tag{28}$$

$\mathbf{R}_a$ and $\mathbf{R}_b$ are defined as follows:

$$\mathbf{R}_a = \mathbf{U}\mathbf{D}\mathbf{V}^T;\ \mathbf{R}_b = \mathbf{U}\mathbf{D}^T\mathbf{V}^T \tag{29}$$

with

$$\mathbf{D} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

This four-fold ambiguity can be solved by using the cheirality constraint, which states that the observed feature points have to be located in front of both cameras. For this it is necessary to reconstruct the three-dimensional coordinates of at least one feature point by using standard triangulation methods and the four possible solutions for the motion parameters. Only in one of those cases the reconstructed point lies in front of both cameras which means that the z-coordinate is bigger than zero.
[42] suggested a more efficient method to test the cheirality constraint which just uses one triangulation and subsequent testing of additional properties which can lead directly to the correct configuration.

4) Guided-RanSaC for handling outliers

Usually the feature detection and matching routine will provide more than five corresponding points between two successive frames of the image sequence. However, it is very likely that the set of point matches contains also a non negligible number of wrong matches (outliers). So there is the open question of choosing the optimal point correspondences for the relative pose estimation.
Thus in literature the calculation of the essential matrix is realised by following Random Sample Consensus (RanSaC) which randomly selects a minimal subset of the data (here: five point matches) and generates an estimate for $\mathbf{E}$ based on those points. Finally all other points are tested against the actual estimation of the essential matrix (e.g. by checking the simplified epipolar constraint from Equation 15). If a sufficient number of point matches are following the estimated model it is assumed to be a correct estimate, otherwise the next minimal subset of points is sampled and the procedure starts again.
Due to the iterative character of the RanSaC approach its usage is neglected within this framework. Instead of a random sampling which treats all samples equally a guided sampling based on a-priori known measures from the feature detection and matching procedure is used here. Similar ideas are described by [38] and [56] within their GOODSaC and GuiSaC procedures.
Most feature detection methods lead to a score which can be interpreted as kind of a distinctiveness measure[6] $\boldsymbol{\xi}$ and also the matching procedure leads to a similarity measure $\boldsymbol{\rho}$. For the experiments incorporating Harris features the distinctiveness $V_{[u,v]}$ at the corner positions defines $\boldsymbol{\xi}$. These information sources are weighted by factors w_ξ and w_ρ to compute an indicator $\boldsymbol{\tau}$ which can be interpreted as the likelihood for being a correct or wrong match.
For the estimation of $\mathbf{E}$ at least five matches are necessary. Hence, the minimal sample sets (MSSs) consist of five matches which are sampled from the set of matches presorted with respect to $\boldsymbol{\tau}$. An iterative procedure is generating estimates for $\mathbf{E}$ by Nisters five-point algorithm until a test of the actual configuration produces a Sampson error d_e over all matches ℓ below a specified threshold d_{lim}. The definition of d_e can be found in Equation 25. Besides that, the number of inliers produced by the actual configuration of $\mathbf{E}$ is evaluated for the stop criterion. The whole procedure for estimating relative camera pose can be described by the following Algorithm 1.
The whole routine is used to generate an estimate for $\mathbf{E}$, whereat it is of course also necessary to apply the concept of the guided-RANSAC scheme for handling the outliers in the set of matched point features. Furthermore the rotation matrix $\mathbf{R}$ and the translation $\mathbf{t}$ are extracted by SVD and subsequent evaluation of the cheirality constraint. The arbitrary scale of $\mathbf{t}$ is determined by incorporating the assumption for t_{init} as the translational movement during the acqui-

[6] It should be stated that the general term *distinctiveness* describes different properties for different feature detectors. So the distinctiveness for a corner-detector would be labelled more exactly as "cornerness" while the features extracted by Fast-Radial Symmetry Transform (FRST) (see [54]) are selected based on their "roundness".

Algorithm 1 Guided-RanSaC procedure for camera egomotion estimation

1: Detect n features in $\mathbf{I}$ and m features in $\mathbf{I}'$ and compute $\boldsymbol{\xi}_i : i \in \{1...n\}$ and $\boldsymbol{\xi}'_j : j \in \{1...m\}$
2: Find ℓ corresponding points $\mathbf{q}_k$ and $\mathbf{q}'_k$ and compute ρ_k with $k \in \{1...\ell\}$
3: **for** all found matches ℓ **do**
4: {Calculate likelihood for being a correct match}
5: $\boldsymbol{\tau}_k = w_\xi \boldsymbol{\xi}_k + w_\rho \boldsymbol{\rho}_k$
6: **end for**
7: Sort all found matches $\mathbf{x}$ and $\mathbf{x}'$ by $\boldsymbol{\tau}$
8: Transform $\mathbf{x}$ and $\mathbf{x}'$ to normalised coordinates $\mathbf{q}$ and $\mathbf{q}'$
9: Sample N MSSs from sorted matches
10: **while** $(d_e < d_{lim}) \wedge (g \leq N) \wedge (h > h_{lim})$ **do**
11: Estimate $\mathbf{E}$ with MSS $g : g \in \{1...N\}$
12: Calculate d_e over ℓ matches
13: Calculate number of inliers h with actual $\mathbf{E}$
14: **end while**
15: Extract $\mathbf{R}_a$, $\mathbf{R}_b$ and $\mathbf{t}_a$, $\mathbf{t}_b$ from $\mathbf{E}$ by SVD
16: Chose correct solution for $\mathbf{R}$ and $\mathbf{t}$ by cheirality constraint

sition of the initial sequence. The following figure gives an overview about the whole procedure, where $\mathbf{x}_{Qj-k}$ describe the matched 2D feature point coordinates in $\mathbf{Q}_j$ and $\mathbf{Q}_k$.

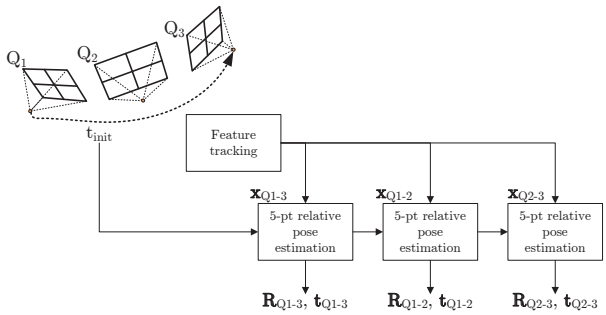


Figure. 15: Relative pose estimation based on three keyframes

5) Preliminary Stereo Triangulation

The generated estimates for $\mathbf{R}$ and $\mathbf{t}$ are used subsequently to determine the preliminary scene model. For this the observed point features which were successfully tracked during the acquisition of the initial sequence, are reconstructed in 3D by standard triangulation techniques (see e.g. [26]). Due to the fact that the translation $\mathbf{t}$ can only be recovered up to an arbitrary scale by Nisters algorithm and the used procedure, which involves the usage of t_{init} is only an assumption about the translational motion between the first and the last frame, the unknown scale between the different two-frame reconstructions has to be resolved. For this the procedure of [27] was used to estimate the scale $\mathbf{s}$ by minimising the term shown in Equation 30, where ${}^{C}\mathbf{X}_{\mathbf{i}}^{\mathbf{Q_{j-k}}} = \begin{bmatrix} x_i & y_i & z_i \end{bmatrix}^T$ describes the 3D reconstruction of the i-th feature point found in both keyframes $\mathbf{Q_j}$ and $\mathbf{Q_k}$. The minimisation of Equation 30 is realised in a least-squares sense.

$$\sum_i \left({}^{C}\mathbf{X}_i^{Q_{1-2}} - \mathbf{s} \cdot {}^{C}\mathbf{X}_i^{Q_{1-3}} \right) \tag{30}$$

6) Optimisation of initial scene model

The initial reconstruction of the scene structure is used as a base for a further refinement by using classical Bundle Adjustment (BA). BA performs a simultaneous optimisation of 3D structure and camera egomotion by minimising the difference between estimated and measured image feature locations ${}^{P}\mathbf{x}_i^k = \begin{bmatrix} u_i^k & v_i^k \end{bmatrix}^T$. In this context the camera or projection matrix of the k-th frame $\mathbf{P}_k$ is used to compute the estimated projections of the 3D structure by following the projection shown in Equation 31, where $\sim$ indicates equality up to scale. Here ${}^{P}\widetilde{\mathbf{x}}_i^k$ describes the i-th 2D point in pixel coordinates for the k-th frame of a sequence in homogenous coordinates. $\mathbf{K}_k$ is the corresponding intrinsic camera matrix and $\mathbf{R}_k$ and $\mathbf{t}_k$ are the corresponding extrinsic parameters for the rigid transformation.

$${}^{P}\widetilde{\mathbf{x}}_i^k \sim \mathbf{K}_k \mathbf{R}_k \left[{}^{C}\widetilde{\mathbf{X}}_i^k - \mathbf{t}_k \right] \tag{31}$$

In general this projection can be formulated by using the projection or camera matrix $\mathbf{P}_k = \mathbf{K}_k \left[\mathbf{R}_k | - \mathbf{t}_k \right]$ as follows:

$${}^{P}\widetilde{\mathbf{x}}_i^k \sim \mathbf{P}_k \, {}^{C}\widetilde{\mathbf{X}}_i^k \tag{32}$$

The procedure of BA consists an interleaving approach based on ideas in [27] and [57] which decouples structure and motion optimisation. The following subsections describe the structure and motion estimation with BA in detail. Besides that it is shown which data is used as initial estimates for both scene structure and camera egomotion, because the provision of adequate initial estimates is crucial for the success of BA-algorithms.

Optimisation of scene structure

The scene structure optimisation is based on the minimisation of the difference between estimated and measured image feature locations. For this the projection in Equation 32 is used as a reference.
The optimisation incorporates all m features, which could be tracked through the whole initialisation sequence with n frames. So the optimal 3D point location for all features can be computed by minimising the following term:

$$\sum_{k=1}^{n} \left[\left(u_i^k - \frac{\mathbf{P}_{k,1}^T \, {}^{C}\widetilde{\mathbf{X}}_i}{\mathbf{P}_{k,3}^T \, {}^{C}\widehat{\mathbf{X}}_i} \right)^2 + \left(v_i^k - \frac{\mathbf{P}_{k,2}^T \, {}^{C}\widetilde{\mathbf{X}}_i}{\mathbf{P}_{k,3}^T \, {}^{C}\widehat{\mathbf{X}}_i} \right)^2 \right] \tag{33}$$

The minimisation is realised in MATLAB by using the Nelder-Mead method as described in [7], where the reconstructed 3D points from the two stereo pairs are used as the initial estimate of scene structure.

Optimisation of camera egomotion

The initial estimates for the camera movement are generated by interpolating the calculated rotations and translations between $\mathbf{Q_1}$ and $\mathbf{Q_2}$, respectively $\mathbf{Q_1}$ and $\mathbf{Q_3}$ from David Nisters algorithm.

The minimisation is based on a nested optimisation procedure which runs one optimisation of scene structure for each iteration of the minimisation of the following error term:

$$\sum_{i=1}^{m} \min_{{}^{C}\widetilde{\mathbf{X}}_i} \left(\sum_{k=1}^{n} \left[\left(u_i^k - \frac{\mathbf{P}_{k,1}^T \, {}^{C}\widetilde{\mathbf{X}}_i}{\mathbf{P}_{k,3}^T \, {}^{C}\widehat{\mathbf{X}}_i} \right)^2 + \left(v_i^k - \frac{\mathbf{P}_{k,2}^T \, {}^{C}\widetilde{\mathbf{X}}_i}{\mathbf{P}_{k,3}^T \, {}^{C}\widehat{\mathbf{X}}_i} \right)^2 \right] \right) \tag{34}$$

It should be stated that it is necessary to update the elements of $\mathbf{P}_k$ for each new iteration of the Nelder-Mead method.
The following figure gives an example for an initial scene model for a planar object (checkerboard on a wall) and the corresponding camera egomotion.

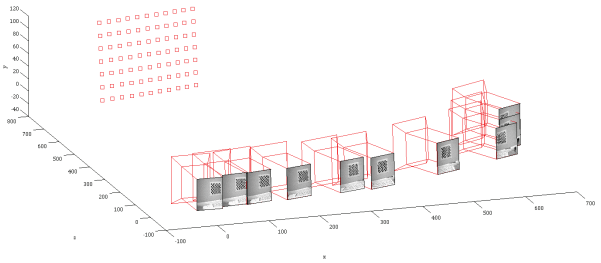


Figure. 16: Example for an initial scene model and the corresponding camera egomotion

B. Sequential SfM

The initial scene model is then used to estimate the camera pose based on 2D/3D correspondences between image features and a scene model which contains calibrated feature positions. For this the 4-point algorithm suggested by [15], because this procedure is also working for a unknown focal length of a camera. This is important to consider if the focal length of the camera can change during the scene acquisition. The following Figure 17 gives an impression about the general configuration of the four-point problem, whereat the relation between object model and image features in the actual frame is visualised.
One major issue in this context is the successful detection and tracking of at least four feature points between the acquired frames and the initial scene model. Of course it is also necessary to add adequate points to the 3D world model to guarantee also the possibility to move around the object which is necessary acquire a complete 3D representation from all sides of the scene. The different parts of the sequential SfM are described in detail in the following paragraphs.

1) Feature detection and matching

The reconstruction of a 3D scene or object from 2D image sequences and the estimation of the camera trajectories are always based on generating correspondences between extracted features from two or more successive frames. For this reason the task can be subdivided into the detection of visual landmarks and their matching in successive frames (image registration) of the sequence. Those features could be of various appearances, whereat most SfM algorithms are based on point features, because the identification of distinctive points

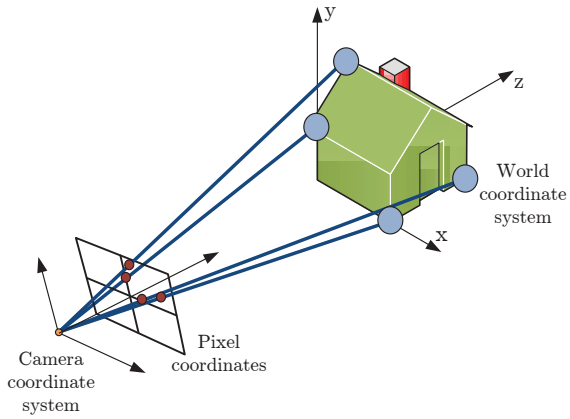


Figure. 17: General configuration of the four-point absolute pose problem

(corners, junctions, etc.) is a well studied field in image processing (see [40]). Also recently published methodologies as SIFT (see [36]) and SURF (see [8]) have drawn the attention of researchers due to their tolerance to scale, illumination and pose variations which can considerably increase the robustness of the registration procedure. For first experiments the Harris corner detector as described by [25] was used for finding distinctive features in the images. The Harris-features are located at the maxima of the local image autocorrelation function $\mathbf{A}$, as shown in the following equation:

$$\mathbf{A} = \begin{bmatrix} \sum\limits_{(i,j)\in\Omega} f_x(i,j)^2 & \sum\limits_{(i,j)\in\Omega} f_x(i,j) \cdot f_y(i,j) \\ \sum\limits_{(i,j)\in\Omega} f_x(i,j) \cdot f_y(i,j) & \sum\limits_{(i,j)\in\Omega} f_y(i,j)^2 \end{bmatrix}$$

The distinctiveness $V_{[u,v]}$ of the points is computed by evaluating $\mathbf{A}$ at image position $[u, v]$ in the following manner:

$$V_{[u,v]} = \det(\mathbf{A}_{[u,v]} - k \cdot trace(\mathbf{A}_{[u,v]})^2 \tag{35}$$

2) Feature tracking

The feature tracking procedure is based on ideas mainly developed by [53] and [31]. Whereat the combination of Markov chains (Kalman filter) and graphical models is used for a robust tracking in 3D. The predicted feature positions are reprojected and based on the reprojected 2D image coordinates a search region for the feature matching procedure can be defined. This is realised by an area-based approach which compares intensity-patches around the feature positions from two successive frames by following the non-parametric ordinal measure as described by [11]. A comparative study of area-based matching techniques in [2] has shown that the ordinal measures outperforms other classical approaches in terms of robustness. Nevertheless there are many possible reasons for wrong matches, as shown e.g. by [6].

3) Absolute pose estimation

As already mentioned above the estimation of the absolute pose is realised by following the algorithm proposed in [15]. The suggested method uses a Groeber basis technique which solves a system of algebraic equations derived from

the number of 2D/3D correspondences generated by the feature tracking and matching routine.

VI. Inertial-Visual Fusion Cell (IVFC)

As mentioned before, one motivation for the implementation of a loosely coupled system as a first stage in the development process is the possibility to run both routes independently. This allows a deep analysis and evaluation of both tracks. By this it will be easily possible to give evidence for the aiding character of the IMU to the SfM-procedure by comparing results of the SfM with and without inference of the inertial track. For the combination of both tracks two unidirectional interfaces will be established between the two routes, as it was shown in Figure 18.
It should be pointed that the visual- and inertial-route will be operating at different frequencies due to the implied sensor devices and the computational elements of both tracks. So for the implementation of the interfaces it is important to consider the multi-rate character of the different measurements. As it was already stated one major problem of the visual measurements is the missing robustness and computational complexity of the feature extraction and tracking. By integrating the pose predictions of the IMU it is possible to considerably limit the search space for feature tracking, because there is an expectation where those features are positioned in the new frame. For this each new feature point inserted into the scene graph is described by a Hidden Markov Model (HMM) (see [20]) for tracking the new feature position based on egomotion estimates from visual and inertial cues.
Moreover it is possible to pre-warp the extracted patches for feature matching based on estimated camera pose. For this the patches are assumed to be locally planar as visualised in Fig. 18. Thus a homography as shown in the Equation 36

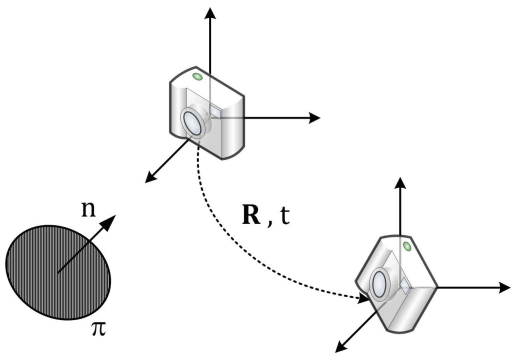


Figure. 18: Locally planar patches for pre-warping

can be computed which relates the patch appearance from one frame to another. Here $\mathbf{K}$ is the intrinsic camera matrix, $\mathbf{R}$ and $\mathbf{t}$ describe the rotation and translation between camera poses, $\mathbf{n}$ is the surface normal, which can be estimated by following the algorithm described in [21], and $\mathbf{x}_p$ is the centre of the patch in the image.

$$\mathbf{H} = \mathbf{K}\mathbf{R}\left[\mathbf{n}^T\mathbf{x}_p\mathbf{I} - \mathbf{t}\mathbf{n}^T\right]\mathbf{K}^{-1} \tag{36}$$

During times of rapid camera movements strength motion blur exists in the imaged frames, so there is the danger that vision-based estimation of egomotion is not possible. In such periods the inertial route would be able to fill the gaps and leading the system while the vision-module is almost not in operation. It was shown by [30] that such a strategy is able to compensate missing measurements from the vision sensor. On the other hand the estimated camera motion from SfM can be used to bound the drift error of the IMU which is a logical consequence from necessary double integration. This is possible due to the generation of position measurements of the visual route and provides a possible solution for an extension of accurate inertial pose predictions for long-term sequences. The realisation of fusing both pose estimates is based on an additional Kalman filter scheme, which incorporates the uncertainties of the separate tracks. It should be pointed out here that the suggested two track system fuses in the Kalman stage pose estimates from both routes which provides the possibility for a relatively simple system and measurement model. It was shown by [17] that it is also possible to fuse the measurements from the MEMS IMU (e.g. 3D acceleration and rotational velocity) directly with pose estimates from the visual route.

VII. Conclusions and future work

The paper proposed a framework for visual-inertial scene reconstruction, whereat the main focus lies on the two distinctive fusion cells for visual and inertial information alone. The actual configuration consists of a parallel fusion network as indicated in Figure 1.
The authors developed also the scheme of a monolithic system design which combines in a single FC the measurements of all four sensory units. The following figure indicates the general architecture of such an entity.

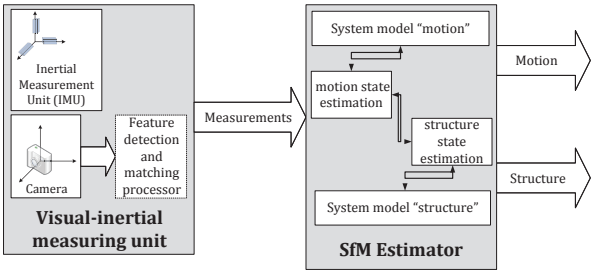


Figure. 19: Monolithic System Design

In such a monolithic or tightly-coupled approach the different sensor units are not longer handled as two separate modules. The camera and the MODS are interpreted as a single visual-inertial sensing device, which provides typical inertial measurements (3D acceleration, angular velocities, earth magnetic field) and feature correspondences as visual measures. Thus the Feature detection and matching processor are formally included in the single measuring unit in this approach. Therefore the definition and implementation of this routine is one major task in this field. Based on the findings from the first stage of the project an enhancement and possible expansion of feature handling is planned at this stage. Furthermore a strategy for handling of multi-rate signals has to be considered and implemented based on the used sensor devices for inertial and visual sensing. Here especially the work of [1], which suggests a multi-rate Unscented Kalman

Filter (UKF) for camera egomotion estimation, shows the potential of multi-rate (MR) sensor fusion.

Acknowledgments

The authors acknowledge support from all the other fellows and students at the Laboratory for Image Processing Soest (LIPS), the Institute for Computer Science, Vision and Computational Intelligence (CV&CI) and all people at the University of Bolton who are involved in this project especially Dr. Dennis Dodds.

References

[1] L. Armesto, S. Chroust, M. Vincze, and J. Tornero, "Multi-rate fusion with vision and inertial sensors," in *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, vol. 1, 2004, pp. 193 – 199 Vol.1.

[2] D. Aufderheide, "Spatial Reconstruction of Head-Geometry by Stereo Vision," Masterthesis, The University of Bolton, 2008.

[3] ——, "VISrec! - Progress reports 2010," Technical report, The University of Bolton, South Westphalia University of Applied Sciences, 2010.

[4] D. Aufderheide and W. Krybus, "A Framework for real time Scene Modeling based on Visual-Inertial Cues," in *IADIS International Conference Computer Graphics, Visualization, Computer Vision and Image Processing 2010 (part of MCCSIS 2010)*, Yingcai Xiao, Tomaz Amon, and Piet Kommers, Eds. IADIS, 2010, pp. 385–390.

[5] ——, "Towards real-time camera egomotion estimation and three-dimensional scene acquisition from monocular image streams," in *2010 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, R. Mautz, M. Kunz, and H. Ingensand, Eds. Zürich, Switzerland: IEEE, 2010, pp. 1–10.

[6] D. Aufderheide, M. Steffens, S. Kieneke, W. Krybus, C. Kohring, and D. Morton, "Detection of salient regions for stereo matching by a probabilistic scene analysis," in *Proceedings of the 9th Conference on Optical 3-D Measurement Techniques*, Wien, 2009, pp. 328–331.

[7] M. Avriel, *Nonlinear Programming: Analysis and Methods*. Dover Publications, 2003.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, 2008.

[9] E. Bayro-Corrochano and B. Rosenhahn, "A geometric approach for the analysis and computation of the intrinsic camera parameters," 2002.

[10] D. Bellot, A. Boyer, and F. Charpillet, "A new definition of qualified gain in a data fusion process: application to telemedicine," in *Information Fusion, 2002. Proceedings of the Fifth International Conference on*, vol. 2, 2002, pp. 865 – 872 vol.2.

[11] D. N. Bhat, D. N. Bhat, S. K. Nayar, and S. K. Nayar, "Ordinal measures for visual correspondence," in *In IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 351–357.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[13] R. S. Blum, *Multi-Sensor Image Fusion and Its Applications (Signal Processing and Communications)*. CRC Press, 2005.

[14] M. Brückner, F. Bajramovic, and J. Denzler, "Experimental Evaluation of Relative Pose Estimation Algorithms," in *VISAPP 2008: Proceedings of the 3rd International Conference on Computer Vision Theory and Applications*, vol. 2, 2008, pp. 431–438.

[15] M. Bujnak, Z. Kukelova, and T. Pajdla, "A general solution to the P4P problem for camera with unknown focal length," in *CVPR 2008, Anchorage, Alaska, USA*, 2008.

[16] M. Caruso, "Applications of magnetic sensors for low cost compass systems," 2000, pp. 177 –184.

[17] S. G. Chroust and M. Vincze, "Fusion of Vision and Inertial Data for Motion and Structure Estimation," *Journal of Robotic Systems*, vol. 21, no. 2, 2004.

[18] P. Corke, J. Lobo, and J. Dias, "An Introduction to Inertial and Visual Sensing," *International Journal of Robotics Research*, vol. 26, no. 6, 2007.

[19] K. Cornelis, M. Pollefeys, M. Vergauwen, L. V. Gool, and K. U. Leuven, "Augmented Reality using Uncalibrated Video Sequences," in *Lecture Notes in Computer Science*, 2001.

[20] A. Davison, *Real-time simultaneous localisation and mapping with a single camera*. IEEE, 2003.

[21] A. Davison, I. Reid, N. Molton, and O. Stasse, "MonoSLAM: real-time single camera SLAM." *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1052–67, 2007.

[22] H. F. Durrant-Whyte, "Sensor Models and Multisensor Integration," *The International Journal of Robotics Research*, vol. 7, no. 6, pp. 97–113, 1988. [Online]. Available: http://ijr.sagepub.com/content/7/6/97.abstract

[23] O. Faugeras and Q.-T. Luong, *The Geometry of Multiple Images: The Laws That Govern the Formation of Multiple Images of a Scene and Some of Their Applications*. The MIT Press, 2004.

[24] D. L. Hall, *Multisensor Data Fusion (Electrical Engineering & Applied Signal Processing Series)*. CRC Press, 2001.

[25] C. Harris and M. Stephens, "A combined corner and edge detection," in *Proceedings of The Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[26] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[27] S. Heung-Yeung, K. Qifa, and Z. Zhengyou, "Efficient bundle adjustment with virtual key frames: a hierarchical approach to multi-frame structure from motion," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. IEEE Comput. Soc, pp. 538–543.

[28] J. Hol, T. Schön, F. Gustafsson, and P. Slycke, "Sensor Fusion for Augmented Reality," in *9th International Conference on Information Fusion*, Florence, Italy, 2006.

[29] J. Hol and P. Slycke, "2D-3D Model Correspondence for Camera Pose Estimation using Sensor Fusion," in *In Proc. of InerVis workshop at the IEEE International Conference on Robotics and Automation*, 2005.

[30] J. D. Hol, "Pose Estimation and Calibration Algorithms for Vision and Inertial Sensors," Masterthesis, Linköping University, 2008.

[31] S. Kieneke, M. Steffens, D. Aufderheide, W. Krybus, C. Kohring, and D. Morton, "Spatio-Temporal Scene Analysis Based on Graph Algorithms to Determine Rigid and Articulated Objects," in *Lecture Notes In Computer Science; Vol. 5496*, 2009.

[32] J. Kim and S. Sukkarieh, "Real-time implementation of airborne inertial-SLAM," *Robotics and Autonomous Systems*, vol. 55, no. 1, pp. 62–71, 2007.

[33] G. Klein and D. Murray, "Parallel tracking and mapping on a camera phone," in *Proc. Eigth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'09)*, Orlando, October 2009.

[34] J. Llinas, *Handbook of Multisensor Data Fusion: Theory and Practice, Second Edition (Electrical Engineering & Applied Signal Processing Series)*. CRC Press, 2008.

[35] J. Lobo, "Inertial Sensor Data Integration in Computer Vision Systems," Masterthesis, University of Coimbra, 2002.

[36] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[37] F. Mata and A. Jimnez, "Multisensor fusion: An autonomous mobile robot," *Journal of Intelligent and Robotic Systems: Theory and Applications*, vol. 22, no. 2, pp. 129–141, 1998.

[38] E. Michaelsen, W. v. Hansen, M. Kirchhof, J. Meidow, and U. Stilla, "Estimating the essential matrix: Goodsac versus ransac," 2006.

[39] H. Mitchell, *Multi-Sensor Data Fusion: An Introduction*. Berlin-Heidelberg: Springer Verlag, 2007.

[40] P. Moreels and P. Perona, "Evaluation of features detectors and descriptors based on 3d objects," *Int. J. Comput. Vision*, vol. 73, no. 3, pp. 263–284, 2007.

[41] T. Morita and T. Kanade, "A sequential factorization method for recovering shape and motion from image streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 8, pp. 858–867, 1997.

[42] D. Nistér, "An efficient solution to the five-point relative pose problem." *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 6, pp. 756–77, 2004.

[43] J.-S. Park, J.-H. Yoon, and C. Kim, "Stable 2D Feature Tracking for Long Video Sequences," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 1, no. 1, pp. 39–46, 2008.

[44] J. Philip, "A Non-Iterative Algorithm for Determining All Essential Matrices Corresponding to Five Point Pairs," *The Photogrammetric Record*, vol. 15, no. 88, pp. 589–599, Oct. 1996.

[45] C. Poelman and T. Kanade, "A paraperspective factorization method for shape and motion recovery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 206–218, 1997.

[46] M. Pollefeys, R. Koch, M. Vergauwen, and L. V. Gool, "Gool. Metric 3D surface reconstruction from uncalibrated image sequences," in *In: 3D Structure from Multiple Images of Large Scale Environments. LNCS Series*. Springer-Verlag, 1998, pp. 138–153.

[47] H. Rehbinder and X. Hu, "Drift-free attitude estimation for accelerated rigid bodies," *Automatica*, vol. 40, no. 4, pp. 653 – 659, 2004.

[48] V. Rodehorst, M. Heinrichs, and O. Hellwich, "Evaluation of Relative Pose Estimation Methods for Multi-camera Setups."

[49] A. Sabatini, "Quaternion-based extended kalman filter for determining orientation by inertial and magnetic sensing," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 7, pp. 1346 –1356, jul. 2006.

[50] J. Sasiadek and Q. Wang, "Sensor fusion based on fuzzy kalman filtering for autonomous robot vehicle," vol. 4, 1999, pp. 2970–2975.

[51] P. G. Savage, "Computational Elements for Strapdown Systems," 2009.

[52] J. Schmidt and H. Niemann, "Using Quaternions for Parametrizing 3-D Rotations in Unconstrained Nonlinear Optimization," in *Vision, Modeling, and Visualization 2001*, T. Ertl, B. Girod, G. Greiner, H. Niemann, and H.-P. Seidel, Eds., Berlin, Amsterdam, 2001, pp. 399–406.

[53] M. Steffens, "Dynamic World Modelling by Dichotomic Information Sets and Graphical Inference with Focus on Facial Pose Tracking," Dissertation, The University of Bolton, 2010.

[54] M. Steffens, S. Kieneke, D. Aufderheide, W. Krybus, C. Kohring, and D. Morton, "Stereo Tracking of Faces for Driver Observation," in *Lecture Notes In Computer Science; Vol. 5575*, 2009.

[55] H. Stewenius, C. Engels, and D. Nister, "Recent developments on direct relative orientation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 60, no. 4, pp. 284–294, Jun. 2006.

[56] B. Tordoff and R. Cipolla, "Uncertain ransac."

[57] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle Adjustment - A Modern Synthesis," *Lecture Notes In Computer Science; Vol. 1883*, 1999.

[58] R. Y. Tsai and T. S. Huang, "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-6, no. 1, pp. 13 –27, 1984.

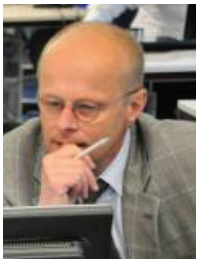
[59] A. Weckenmann, X. Jiang, K.-D. Sommer, U. Neuschaefer-Rube, J. Seewig, L. Shaw, and T. Estler, "Multisensor data fusion in dimensional metrology," *CIRP Annals - Manufacturing Technology*, vol. 58, no. 2, pp. 701 – 721, 2009.

[60] J. Wendel, *Integrierte Navigationssysteme*. Oldenbourg Wissensch.Vlg, 2007.

[61] K. Z., B. M., and P. T., "Polynomial eigenvalue solutions to the 5-pt and 6-pt relative pose problems," in *Proceedings. BMVC 2008)*, 2008.

Author Biographies

Dominik Aufderheide is an active researcher in the area of multi-sensor image processing and computer vision. He studied Electrical Engineering with a focus on embedded systems and signal processing at the South Westphalia University of Applied Sciences in Soest, Germany and graduated 2007 with the German diploma. Afterwards he became part of an international master course at the University of Bolton, U.K. focused on electronic system design and engineering management, where he received 2009 a M.Sc. with Distinction. Currently he is a research fellow at the Institute of Computer Science, Vision and Computational Intelligence (CV&CI) in Soest, where he is working towards a Ph.D. degree in cooperation with the University of Bolton. His research interests are mainly focused on multi-sensor data fusion, computer vision, machine learning and smart sensor systens. Dominik Aufderheide is IEEE student member since 2007.



Werner Krybus is professor at South Westphalia University of Applied Sciences in Soest, Germany. He studied Electrical Engineering at RWTH Aachen University and graduated in 1984 . He received his Ph.D. from RWTH Aachen University on a topic about computer-assisted surgery. In 1996 W. Krybus joined the South Westphalia University of Applied Sciences as a professor for data systems engineering and signal processing. Dr. Krybus is founder of the Laboratory for Image Processing Soest within the Institute for Computer Science, Vision and Computational Intelligence. His primary research interests include embedded systems, signal processing and sensor fusion.