Using Natural Features for Vision Based Navigation of an Indoor-VTOL

Christian Schlaile, Oliver Meister, Jan Wendel, and Gert F. Trommer
Institute of Theory and Systems Optimization in Electrical Engineering (ITE)
http://www.ite.uni-karlsruhe.de
Universität Karlsruhe (TH), Kaiserstr. 12, D-76128 Karlsruhe, Germany
Tel. +49 (0)721/608-2641, Fax. +49 (0)721/608-2623
E-mail: Christian.Schlaile@ite.uni-karlsruhe.de

Abstract

In this paper, the use of natural features for vision based navigation of an Indoor-VTOL named Air-Quad is presented. Air-Quad is a small four-rotor helicopter developed at the ITE.

Such a helicopter needs reliable attitude information. The used MEMS gyroscopes and accelerometers have strong noise and need aiding through complementary sensors like GPS or the here presented computer vision module.

In the computer vision module, feature points are tracked through the image sequence and keyframes are determined. With the two-dimensional coordinates of the feature points the relative rotation and translation of the camera between the keyframes are estimated. The three-dimensional coordinates of points in the scene are triangulated. An efficient sparse bundle adjustment algorithm is used to improve the estimation of the scene structure and the navigation solution.

It is shown that the use of the computer vision module improves greatly the navigation solution compared to a solution based only on MEMS sensors.

As a first step towards collision avoidance, the free space in front of the camera is estimated as a polyhedron approximation.

1 Introduction

In the last years, small Unmanned Aerial Vehicles (UAV) called Micro Aerial Vehicles (MAV) were developed to be used in surveillance and reconnaissance tasks.

Especially for the control of helicopters an accurate attitude information is necessary. The used cheap Micro-Electro-Mechanical Systems (MEMS) gyroscopes and accelerometers are noisy and need additionally aiding sensor information to supply a reliable navigation information. Such an additionally sensor could be GPS. But GPS fails in indoor environments, cannot aid the yaw angle estimating during hovering, or can be jammed. Therefore the use of a computer vision module for the aiding of the navigation system is an interesting option.

Figure 1: Flying airframe Air-Quad with mounted camera at the left side.

Due to the increased availability of computer power, computer vision systems are getting used for MAV in the last few years: A simulation of a GPS/INS system for a fixed wing aircraft with horizon detection is shown in [1]. In [2] the speed over ground is calculated using a stereo camera. In [3] a camera based control of a four-rotor helicopter using colored markers is shown. Artificial markers to determine the position and attitude are presented in [4].

In this contribution, the possibility to use a computer vision module to aid the position and attitude estimation using non-artificial features is shown with real data.

This paper is divided in the following parts: In the next section, the used airframe Air-Quad is presented, after that a description of the integrated navigation systems follows. Then the developed computer vision

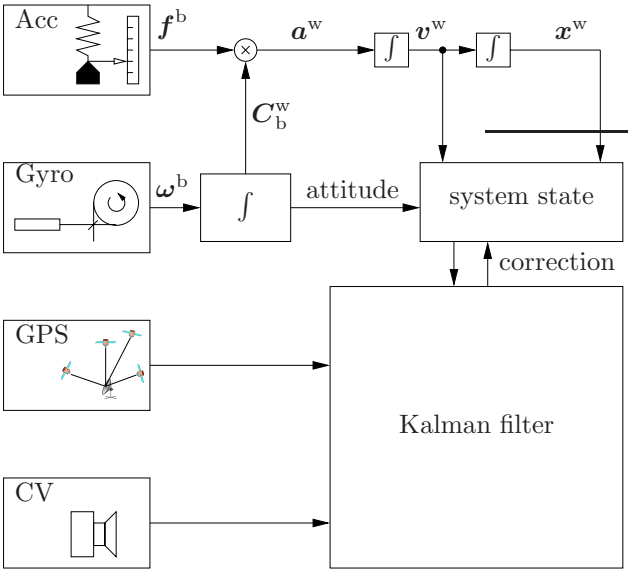

Figure 2: Integrated navigation system.

module is described in great detail. After that, the experimental results are presented. And finally, conclusions are drawn.

2 AIRFRAME AIR-QUAD

A four rotor helicopter named Air-Quad (s. Fig. 1) was designed at the ITE to develop and verify navigation algorithms. Its advantages are the simple mechanical design without swash plates and the high agility. The control of the position, velocity and attitude is achieved using different rotational speeds of the rotors. For example, a roll to the left is made by reducing the speed of the left rotor and increasing the speed of the right rotor.

Reliably navigation information provided by an integrated navigation system is required to control the helicopter.

3 INTEGRATED NAVIGATION SYSTEM

An integrated navigation system (s. Fig. 2) integrates the measured accelerations $\boldsymbol{f}^{\mathrm{b}}$ and rotations $\boldsymbol{\omega}^{\mathrm{b}}$ using the strapdown algorithm [5] to gain attitude $\boldsymbol{C}_{\mathrm{b}}^{\mathrm{w}}$, position $\boldsymbol{x}^{\mathrm{w}}$ and velocity $\boldsymbol{v}^{\mathrm{w}}$ information. Because of the noise of the used MEMS sensors, additionally aiding with complementary sensors like GPS [6] or the here presented computer vision module is essential.

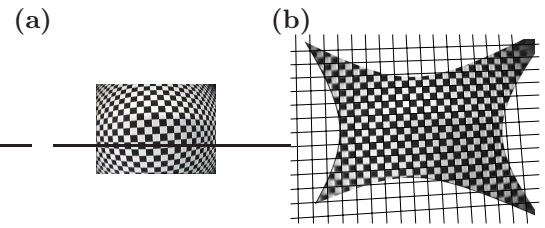

Figure 3: Lens distortion: **(a)** image captured with the wide angle camera, **(b)** removed distortion.

4 COMPUTER VISION MODULE

The task of the computer vision module is to determine the camera movement using the shift of image points.

4.1 Camera model

The camera model describes the relationship between the three-dimensional points in the scene and their two-dimensional image coordinates.

A three-dimensional point $(x, y, z)^{\mathrm{T}}$ can be written in homogenous coordinates as $(x, y, z, 1)^{\mathrm{T}}$. This allows the combination of rotation $\boldsymbol{R} = (r_{ij})$ and translation $\boldsymbol{t} = (t_x, t_y, t_z)$ to the transformation $\mathcal{T}_{\mathrm{a}}^{\mathrm{b}}$. It transforms the vector $\boldsymbol{x}^{\mathrm{a}}$ given in coordinate system a to the vector $\boldsymbol{x}^{\mathrm{b}}$ given in coordinate system b.

$$\boldsymbol{x}^{\mathrm{b}} = \mathcal{T}_{\mathrm{a}}^{\mathrm{b}} \cdot \boldsymbol{x}^{\mathrm{a}} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}. \tag{1}$$

By using the homogenous coordinates the equal sign "=" means equality only up to a scale. The projection of a three-dimensional vector given in the camera frame to the two-dimensional image coordinates $(u, v)^{\mathrm{T}}$ is [7]:

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} f_x & 0 & p_x \\ 0 & f_y & p_y \\ 0 & 0 & 1 \end{pmatrix}}_{\boldsymbol{K}} \cdot [\boldsymbol{I}|0] \cdot \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{2}$$

The matrix $\boldsymbol{K}$ is the camera calibration matrix, the principal point is $(p_x, p_y)^{\mathrm{T}}$, the focal lengths are (f_x, f_y), and $\boldsymbol{I}$ is the 3×3-identity matrix. The matrix $\boldsymbol{K}$ and the parameter of the lens distortion (s. Fig. 3) of the used wide angle camera are determined in a preflight calibration.

4.2 Feature points selection and tracking

Feature points are found in the real images using the Harris corner detector [8]. In the adjacent images of

the video sequence the feature points are tracked using the optical flow. A stable feature point can be tracked from the first image to the second image and back to the first image resulting in the same image coordinates. Noise and mismatches are common, so all used models are robustly estimated using RANdom SAmple Consensus (RANSAC)[9]. These models are described in the following.

4.3 Point movement models

Two models describing the point movement in an image sequence are presented in the following two subsections.

Fundamental matrix

With a camera translation $\boldsymbol{t}$ between two camera views $\boldsymbol{P}_1 = \boldsymbol{K} \cdot [\boldsymbol{I}|\boldsymbol{0}]$ and $\boldsymbol{P}_2 = \boldsymbol{K} \cdot [\boldsymbol{R}|\boldsymbol{t}]$ the three-dimensional point $\boldsymbol{X}$ are projected to $\boldsymbol{x}_1 = \boldsymbol{K} \cdot [\boldsymbol{I}|\boldsymbol{0}] \cdot \boldsymbol{X}$ and $\boldsymbol{x}_2 = \boldsymbol{K} \cdot [\boldsymbol{R}|\boldsymbol{t}] \cdot \boldsymbol{X}$.

The matrix $\boldsymbol{F}$ with $\boldsymbol{x}_2^{\mathrm{T}} \cdot \boldsymbol{F} \cdot \boldsymbol{x}_1 = 0$ is called fundamental matrix [10]. It is

$$\boldsymbol{F} = \boldsymbol{K}^{-\mathrm{T}} \cdot [\boldsymbol{t}]_\times \cdot \boldsymbol{R} \cdot \boldsymbol{K}^{-1}, \tag{3}$$

with $[\boldsymbol{t}]_\times \cdot \boldsymbol{a} = \boldsymbol{t} \times \boldsymbol{a}$.

With a translation $\boldsymbol{t} \neq \boldsymbol{0}$, the relative rotation and translation of the two cameras can be calculated using the fundamental matrix. With the found rotation and translation of the camera and matched feature points, the coordinates of the three-dimensional points can be calculated using triangulation.

With no or a very small translation and noisy point coordinates the determination of the fundamental matrix is ill-conditioned. In this case, it is not possible to gain information about the rotation.

The essential matrix $\boldsymbol{E} = [\boldsymbol{t}]_\times \cdot \boldsymbol{R}$ could be used instead of $\boldsymbol{F}$ if the camera calibration matrix $\boldsymbol{K}$ is known.

Homography

A homography is the mapping of points from one plane (e.g. the first image plane) to points on a second plane (e.g. the second image plane or a planar part of the scene).

A homography describes the scene perfectly, if the camera only rotates between the used images. Having only a pure rotation, it is not possible to gain the depth of the points.

And a homography describes a pure planar scene perfectly. With given camera calibration, the homography [7] for the cameras $\boldsymbol{P}_1 = \boldsymbol{K} \cdot [\boldsymbol{I}|\boldsymbol{0}]$ and $\boldsymbol{P}_2 = \boldsymbol{K} \cdot [\boldsymbol{R}|\boldsymbol{t}]$ is

$$\boldsymbol{H} = \boldsymbol{R} - \boldsymbol{t} \cdot \boldsymbol{n}^{\mathrm{T}}/d \qquad \text{with} \qquad \boldsymbol{x}_2 = \boldsymbol{H} \cdot \boldsymbol{x}_1 . \tag{4}$$

Here $\boldsymbol{\pi}_E = (\boldsymbol{n}^{\mathrm{T}}, d)^{\mathrm{T}}$ describes the seen plane.

With a pure rotation ($\boldsymbol{t} = \boldsymbol{0}$) the rotation matrix can be taken directly from the estimated homography.

4.4 Selecting Model and Keyframes

The fundamental matrix and homography are both useful to find new point correspondences. The homography can be used to describe pure rotation or pure planar scene. The fundamental matrix describes a general scene if there is a translation and can provided depth information via triangulation. Next, the selection of the best fitting model is described.

Model selection using GRIC

The geometric robust information criterion GRIC is suggested in [11] to choose between different models. It consists of two parts: One part for the parsimony of the model and one part for the goodness of the fit. For every movement model the unit-less GRIC is calculated:

$$\mathrm{GRIC} = \sum_i \rho(e_i^2) + d \cdot n \cdot \ln(r) + k \cdot \ln(r \cdot n) . \tag{5}$$

Where n is the count of features, e_i are the residues, r is the dimension of the data (4 for two views), k is the dimension of the model parameter (7 for $\boldsymbol{F}$ and 8 for $\boldsymbol{H}$), and d is the structure dimension (3 for $\boldsymbol{F}$ and 2 for $\boldsymbol{H}$).

The robust function $\rho(e^2)$ of the residues of the model fitting is

$$\rho(e^2) = \min\left(\frac{e^2}{\sigma^2}, 2 \cdot (r - d)\right) \tag{6}$$

with σ as the standard deviation of the measurement errors.

The different models could be compared directly using GRIC: The model with the lowest GRIC value describes the point shift best.

Choice and use of keyframes

The homography is usually the best model for two adjacent images of a video sequence. To get information about the depth of the scene, it is necessary to find two camera views with enough translation between them.

To achieve this goal, keyframes are selected from the sequence of the images. The first image is always the keyframe S_0. The next keyframes are chosen according to the following procedure: The fundamental matrix and the homography and their GRIC are robustly estimated between the current image and the current keyframe S_i. A new keyframe is found if GRIC-E falls below GRIC-H (s. Fig. 4).

In the following, the coordinate systems at the time of the preceding keyframe S_{i-1} have the suffix 0 and

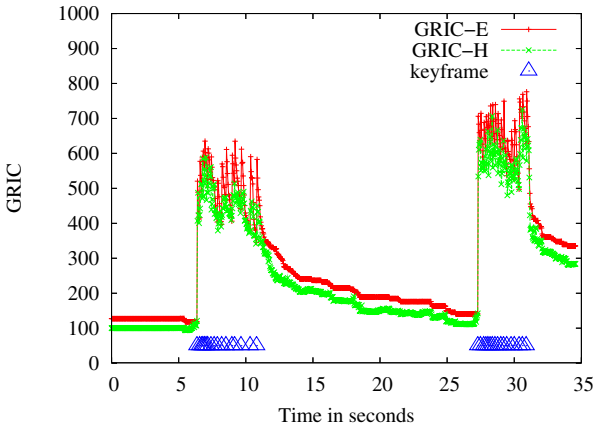

Figure 4: Found keyframes (△) and geometric robust information criterion for the homography $\boldsymbol{H}$ (GRIC-H) and essential matrix $\boldsymbol{E}$ (GRIC-E) calculated from current frame to the preceding keyframe.

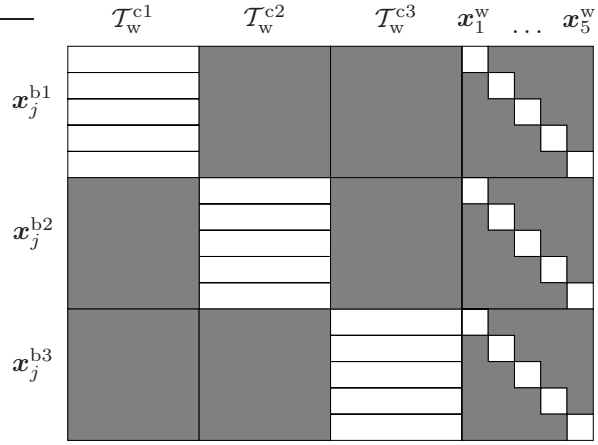

Figure 5: Sparse Jacobian matrix for the bundle adjustment of three cameras and five features: Gray fields are zero.

the coordinate systems at the time of the current image have the suffix 1. The attitude and position of the strapdown algorithm at the time of the preceding keyframe S_{i-1} have been stored as $\mathcal{T}_{\mathrm{b0}}^{\mathrm{w}}$.

The rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{t}$ (up to an unknown scale) between the current image and the preceding keyframe S_{i-1} can be determined using the found essential matrix $\boldsymbol{E} = [\boldsymbol{t}]_\times \cdot \boldsymbol{R}$ according to [7] and results in the transformation $\mathcal{T}_{\mathrm{c1}}^{\mathrm{c0}}$.

4.5 Triangulation and Bundle Adjustment

With the point correspondences between the current keyframe S_i and the preceding keyframe S_{i-1} three-dimensional points of the scene can be triangulated and added into a sparse map of the environment using the estimated rotation $\boldsymbol{R}$ and translation $\boldsymbol{t}$ between the two keyframes.

The three-dimensional point $\boldsymbol{X}$ given in the camera frame of the keyframe S_{i-1} is projected with $\boldsymbol{P}_1 = \boldsymbol{K} \cdot [\boldsymbol{I}|0]$ and $\boldsymbol{P}_2 = \boldsymbol{K} \cdot [\boldsymbol{R}|\boldsymbol{t}]$ to the image coordinates $(u_1, v_1)^{\mathrm{T}}$ and $(u_2, v_2)^{\mathrm{T}}$. The equation $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}$ with

$$\boldsymbol{A} = \begin{pmatrix} u_1 \cdot \boldsymbol{p}_1^{3\mathrm{T}} - \boldsymbol{p}_1^{1\mathrm{T}} \\ v_1 \cdot \boldsymbol{p}_1^{3\mathrm{T}} - \boldsymbol{p}_1^{2\mathrm{T}} \\ u_2 \cdot \boldsymbol{p}_2^{3\mathrm{T}} - \boldsymbol{p}_2^{1\mathrm{T}} \\ v_2 \cdot \boldsymbol{p}_2^{3\mathrm{T}} - \boldsymbol{p}_2^{2\mathrm{T}} \end{pmatrix} \tag{7}$$

and the rows of $\boldsymbol{P}_j$ given as $\boldsymbol{p}_j^{i\mathrm{T}}$ can be solved for $\boldsymbol{X}$. The found three-dimensional points and camera frames are used as initial parameters for an iterative Levenberg-Marquardt optimization of the reprojection error for all camera frames and points. The efficient implementation of this bundle adjustment takes advantages of the sparse nature (s. Fig. 5) of the Jacobian matrix [7] of the reprojection error.

The bundle adjustment allows the optimal estimation of points in the scene and all camera poses (translation and rotation). It is also possible to quickly estimate only the current camera pose by keeping the scene structure and the old camera frames constant. Using only the computer vision module, it is possible with one camera to determine scene and translation up to a global scale (doll house effect). With additional sensor data (e.g. IMU) this scale factor can be determined.

The result of the bundle adjustment for the current camera frame $\tilde{\mathcal{T}}_{\mathrm{c1}}^{\mathrm{w}}$ is compared to the transformation $\mathcal{T}_{\mathrm{b1}}^{\mathrm{w}}$ taken from the strapdown algorithm under the assumption that the transformation $\mathcal{T}_{\mathrm{c1}}^{\mathrm{b1}}$ between the body frame b and camera frame c is constant.

The result of this comparison can be used in a Kalman filter to improve the state of the strapdown algorithm. The measurement of the Kalman filter is:

$$\boldsymbol{\Delta} = \tilde{\mathcal{T}}_{\mathrm{b1}}^{\mathrm{w}} \cdot \mathcal{T}_{\mathrm{w}}^{\mathrm{b1}} \,. \tag{8}$$

The angles in $\boldsymbol{\Delta}$ are small enough to use an Euler angles formulation.

5 EXPERIMENTAL RESULTS

Three trajectories are used for the evaluation of the computer vision module with real video data: The first and second trajectories are used to estimate the accuracy of the computer vision module. The last trajectory is used to measure the speed of the algorithm and perform first steps towards collision avoidance.

Accuracy Estimation

The first trajectory is a fast raise from the ground level followed by a decline with steps of equal size. The result of the computer vision module is shown in Fig. 6

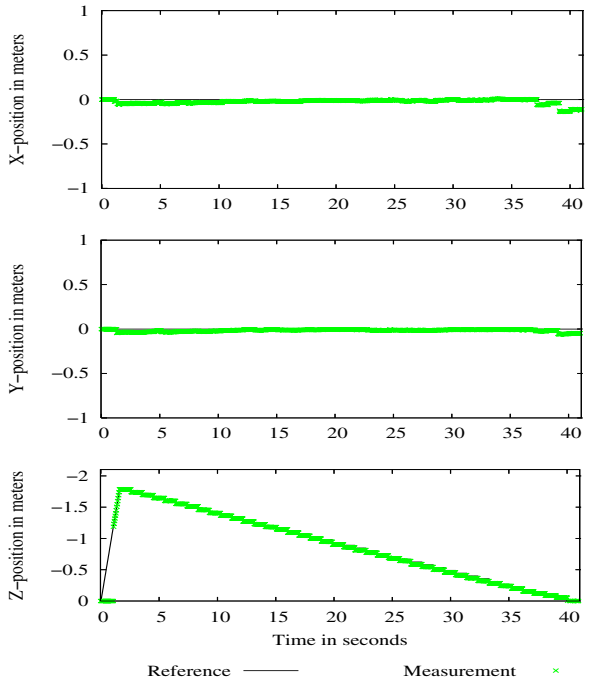

Figure 6: Position measurements generated by the computer vision module in comparison with the reference.

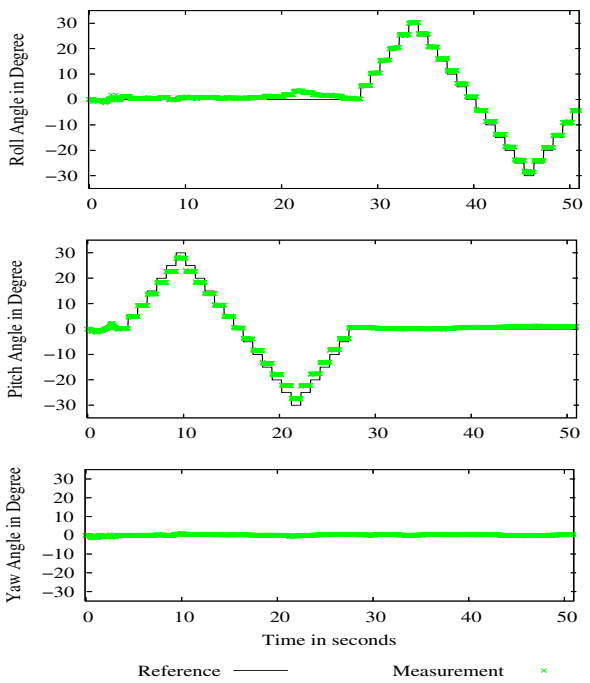

Figure 7: Attitude measurements as Euler angles generated by the computer vision module in comparison with the reference.

together with the reference. In the beginning of the sequence, the bundle adjustment needs some movement of the camera to get a first estimation of the scene structure. A mean absolute error of $(2.4, 1.6, 3.7)$ cm is achieved with a maximum flight height of 178.4 cm.

For the second trajectory, a two-axis table with angle scale is used to determine the reference of the Euler angles (roll, pitch, yaw). The result is shown in Fig. 7. A mean absolute Euler angles error of $(0.91, 1.02, 0.28)$ deg is achieved. This error is drift-free and enables a control of the hover flight.

Without the aiding of the computer vision module, the MEMS based solution drifts in 35 seconds to a position error greater than 500 m and an attitude error for the yaw-angle greater than 7 deg. Using computer vision, even the yaw angle could be stabilized in a hover flight, which is not possible using a GPS receiver.

Real-Time Capability

The third trajectory starts at the ground level of a corridor. The hand-carried camera hovers and turns slightly around the yaw-axis. In the end, it landed on the starting spot.

Using a desktop PC, the computer vision module has the following timing: The first bundle adjustment using 12 keyframes and 525 image projections needs 260 ms. Adding one keyframe is done in 25.2 ms and estimating the current camera frame without optimization the scene structure needs 0.5 ms. With a recorded video sequence, it is possible to process 28 frames per second. This result permits the aiding of the strapdown algorithm in real-time.

Collision Avoidance

A first step towards collision avoidance is shown in Fig. 8: Starting with the two-dimensional positions of feature points in the first image, a Delaunay subdivision is calculated to gain a triangular mesh. This mesh is transformed into a three-dimensional approximation of the seen planar surfaces using the estimated three-dimensional position of the used points. The outer edges of the mesh are connected with the camera center to form a polyhedron of free space. The combination of polyhedrons from different camera view points could be used in the future to build a collision avoidance system for indoor flying helicopters.

6 CONCLUSIONS

In this paper, the use of natural features for vision based navigation of the four-rotor helicopter Air-Quad was presented.

Two models to describe the shift of features were investigated using real image data: The homography which describes a pure rotation or a planar scene and the fundamental matrix which describes a general scene under the condition that a translation exists.

The models were evaluated using a geometric robust information criterion. Keyframes were selected based

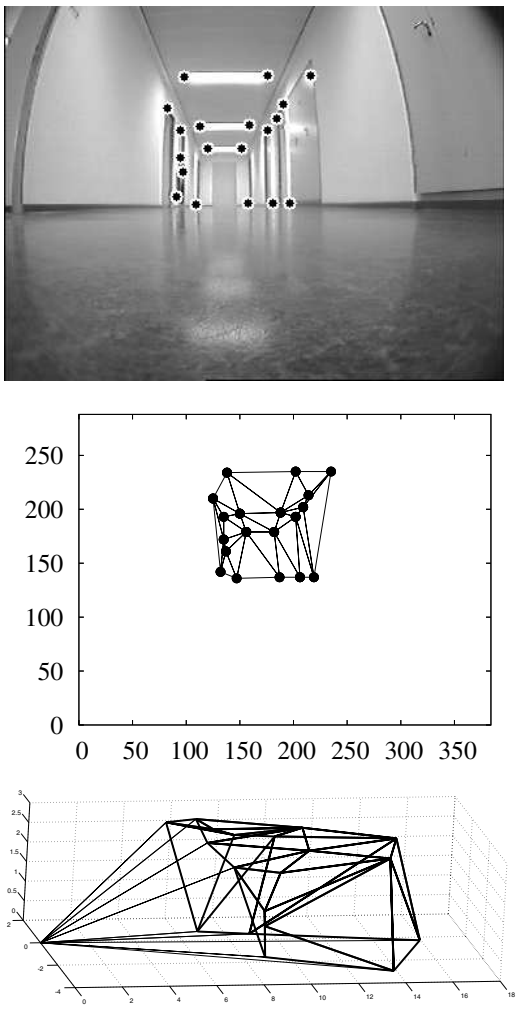

Figure 8: Feature points in real video data, Delaunay subdivision, resulting polyhedron of free space.

on this criterion.

Feature point coordinates in the keyframes were given to an efficient sparse bundle adjustment algorithm to optimize the estimation of the structure of the scene and the camera position and attitude.

The results demonstrated the use of a computer vision module without artificial landmarks to aid a navigation system in an indoor environment. A drift-free navigation solution for position and attitude (important for the hover flight) could be achieved. In opposite to a GPS based solution, even the yaw angle could be stabilized using the computer vision module while hovering.

First steps towards collision avoidance results in the estimation of a polyhedron of free space in front of the camera.

The next steps will be the use of the found free space to build a collision avoidance system for indoor flying helicopters.

References

[1] S. Winkler, H.-W. Schulz, M. Buschmann, T. Kordes, and P. Vörsmann. Improving low-cost GPS / MEMS-based INS Integration for Autonomous MAV Navigation by Visual Aiding. In *Proceedings of the ION GNSS 2004*, Long Beach, CA, 21-24 Sept. 2004.

[2] Omead Amidi. *An Autonomous Vision-Guided Helicopter*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 1996.

[3] E. Altug, J.P. Ostrowski, and C.J. Taylor. Quadrotor control using dual camera visual feedback. In *IEEE International Conference on Robotics and Automation*, volume 3, pages 4294–4299, 2003.

[4] Christian Schlaile, Jan Wendel, and Gert F. Trommer. Stabilizing a Four-Rotor Helicopter using Computer Vision. In *Proceedings of the 1st European Micro Air Vehicle Conference and Flight Competition (EMAV2004)*, Braunschweig, Germany, 2004.

[5] J.L. Titterton, D.H.; Weston. *Strapdown inertial navigation technology*. Peter Peregrinus Ltd./IEE, London, 1997.

[6] Jan Wendel, Christian Schlaile, and Gert F. Trommer. Integration of Low-Cost MEMS Inertial Sensors and GPS for Unmanned Aerial Vehicles. Dresden, 20. bis 23. September 2004. Deutscher Luft- und Raumfahrtkongress.

[7] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge Univ. Press, second edition, 2003.

[8] Chris Harris and Mike Stephens. A Combined Corner and Edge Detector. In *4th Alvey Conference*, pages 147–151, Manchester, UK, August 1988.

[9] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[10] Olivier Faugeras and Quang-Tuan Luong. *The geometry of multiple images : The Laws That Govern the Formation of Multiple Images of a Scene and Some of Their Applications*. MIT Press, 2001.

[11] Philip H. S. Torr, Andrew W. Fitzgibbon, and Andrew Zisserman. The problem of degeneracy in structure and motion recovery from uncalibrated image sequences. *Int. J. Comput. Vision*, 32(1):27–44, 1999.