Linköping Studies in Science and Technology
Thesis No. 1672

Global Pose Estimation from Aerial Images

Registration with Elevation Models

Bertil Grelsson

Linköping University
INSTITUTE OF TECHNOLOGY

Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden

Linköping June 2014

Global Pose Estimation from Aerial Images

© 2014 Bertil Grelsson

Cover image: Fisheye image from Östergötland, Sweden, with estimated horizon overlaid. 3D model of Rio de Janeiro based on satellite imagery, courtesy of SAAB Vricon Systems AB.

Department of Electrical Engineering
Linköping University
SE-581 83 Linköping
Sweden

ISBN 978-91-7519-279-6 ISSN 0280-7971

Abstract

Over the last decade, the use of unmanned aerial vehicles (UAVs) has increased drastically. Originally, the use of these aircraft was mainly military, but today many civil applications have emerged. UAVs are frequently the preferred choice for surveillance missions in disaster areas, after earthquakes or hurricanes, and in hazardous environments, e.g. for detection of nuclear radiation. The UAVs employed in these missions are often relatively small in size which implies payload restrictions.

For navigation of the UAVs, continuous global pose (position and attitude) estimation is mandatory. Cameras can be fabricated both small in size and light in weight. This makes vision-based methods well suited for pose estimation onboard these vehicles. It is obvious that no single method can be used for pose estimation in all different phases throughout a flight. The image content will be very different on the runway, during ascent, during flight at low or high altitude, above urban or rural areas, etc. In total, a multitude of pose estimation methods is required to handle all these situations. Over the years, a large number of vision-based pose estimation methods for aerial images have been developed. But there are still open research areas within this field, e.g. the use of omnidirectional images for pose estimation is relatively unexplored.

The contributions of this thesis are three vision-based methods for global ego-positioning and/or attitude estimation from aerial images. The first method for full 6DoF (degrees of freedom) pose estimation is based on registration of local height information with a geo-referenced 3D model. A dense local height map is computed using motion stereo. A pose estimate from navigation sensors is used as an initialization. The global pose is inferred from the 3D similarity transform between the local height map and the 3D model. Aligning height information is assumed to be more robust to season variations than feature matching in a single-view based approach.

The second contribution is a method for attitude (pitch and roll angle) estimation via horizon detection. It is one of only a few methods in the literature that use an omnidirectional (fisheye) camera for horizon detection in aerial images. The method is based on edge detection and a probabilistic Hough voting scheme. In a flight scenario, there is often some knowledge on the probability density for the altitude and the attitude angles. The proposed method allows this prior information to be used to make the attitude estimation more robust.

The third contribution is a further development of method two. It is the very first method presented where the attitude estimates from the detected horizon in omnidirectional images is refined through registration with the geometrically expected horizon from a digital elevation model. It is one of few methods where the ray refraction in the atmosphere is taken into account, which contributes to the highly accurate pose estimates. The attitude errors obtained are about one order of magnitude smaller than for any previous vision-based method for attitude estimation from horizon detection in aerial images.

Acknowledgements

First of all, I would like to express my sincere gratitude to all former and current members of the Computer Vision Laboratory for providing an inspiring and friendly working environment. Although being an outlier from industry, I have really felt like a member of the group. I would especially like to thank

- My supervisor Michael Felsberg for excellent guidance and support and being a true source of inspiration, always perceiving opportunities in the technical challenges.

- My co-supervisor Per-Erik Forssén for fruitful detailed discussions on my research topic and for giving me valuable insights how to compose and write scientific papers.

- The CIMSMAP project group comprising Michael Felsberg, Per-Erik Forssén, Leif Haglund, Folke Isaksson, Sören Molander and Pelle Carlbom for providing great technical support and advice, and at each meeting generating a diversity of conceivable research paths, some of them leading to this thesis, some of them still being unexplored.

I would also like to thank my family, colleagues and friends for their everyday life support, most notably

- Annika and Hanna for your love, patience and understanding during this period. Your large share of family chores is gratefully appreciated. Assisting me with the figures was truly helpful when writing this thesis. And, most importantly, thanks for now and then reminding me that there is more in life than computations on images. An early morning in the stables can be a true life recharger.

Thanks also to my employer Saab Dynamics for giving me the opportunity to undertake the studies leading to this thesis. Hopefully it will turn out a joint win-win situation in the long run.

This work was funded by the Swedish Governmental Agency for Innovation Systems, VINNOVA, under contracts NFFP5 2010-01249 and 2013-05243.

Bertil Grelsson June 2014

Contents

Part I

Background Theory

Chapter 1

Introduction

1.1 Motivation

When pilots navigate an aircraft to its destination they use information provided to them originating from instrumentations like radars, radio beacons, GPS, inertial sensors, altimeters, compasses, etc., see figure 1.1. But, especially for less equipped aircraft, pilots also use a lot of visual cues for navigation during the flight. Landmarks such as buildings, lakes, roads, mountains in the terrain are used for coarse global positioning of the aircraft. The aircraft heading can be deduced from the relative position of the landmarks in the pilot's view when comparing it with map information. Furthermore, the horizon line is often used as an attitude guide during takeoff and landing.

In recent years, the use of unmanned aerial vehicles (UAVs) has increased drastically. The flight of a UAV is controlled either autonomously by computers onboard the vehicle or under remote control by a pilot on the ground. Originally, the use of UAVs was mainly military, but nowadays many civil applications have emerged. UAVs may e.g. be used for surveillance of pipelines or power lanes. They are also often preferred for missions in hazardous environments for detection of nuclear radiation or in disaster areas after earthquakes or hurricanes.

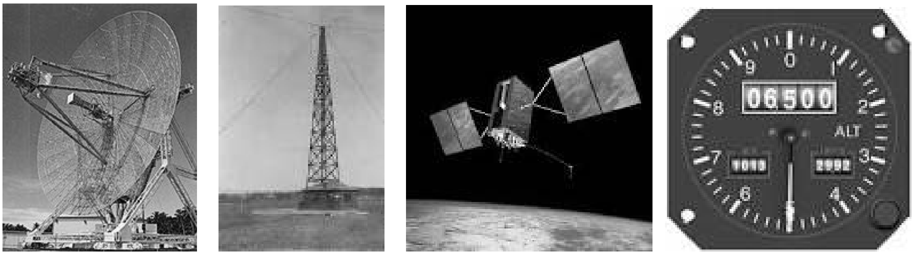

Figure 1.1: Navigation aids. From left to right: radar antenna, radio beacon, GPS satellite, altimeter display.

For navigation of the UAVs, continuous global pose (position and attitude) estimation is mandatory. The UAVs employed in these missions are often relatively small in size which implies payload restrictions, both in size and weight. This is where the characteristics of vision-based methods for ego-positioning and navigation of the UAVs are appealing. Cameras can be fabricated both small in size and light in weight. A very large amount of image data can be captured and stored onboard the vehicle for processing post flight. The processing required for global pose estimation and navigation can also often be performed in real time onboard the vehicle. The information viewed in the images can thus be georeferenced which makes vision-based methods an ideal fit for the applications mentioned above.

1.2 Goals of this thesis

The aim of the work leading to this thesis was to develop automated vision-based methods for global pose (position and orientation) estimation of airborne vehicles, see figure 1.2. Global, in this context, means that the pose is given in a world coordinate frame, relative to the earth. The research work has been conducted within the framework of a project called CIMSMAP, "Correlation of image sequences with 3D mapping data for positioning in airborne applications", a collaboration between Linköping University and SAAB. The main idea was to perform global pose estimation via registration of aerial images with a geo-referenced 3D model generated from aerial images captured in a previous instance in time. The developed methods for global pose estimation have been evaluated using true aerial imagery captured in flights with manned aircraft. Example images, used for pose estimation, from these flights are shown in figure 1.2.

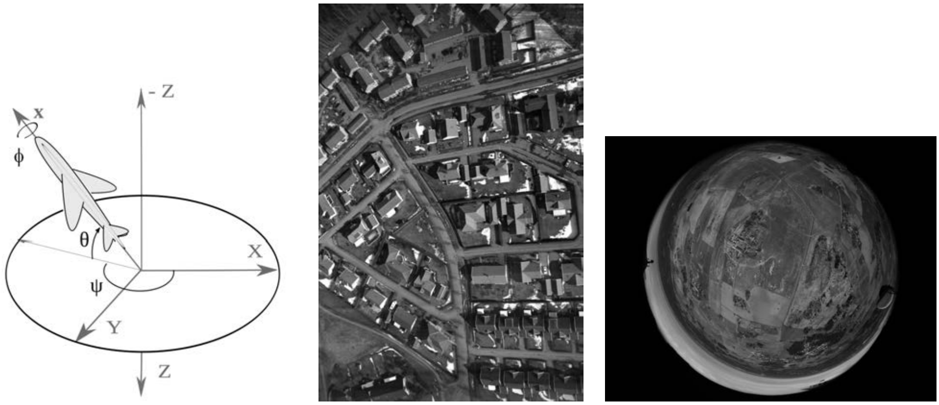

Figure 1.2: Definition of vehicle pose in world coordinate frame (left). X = north, Y = east, Z = down, ψ = yaw, θ = pitch, ϕ = roll. Example images from flights used in paper A (middle) and paper C (right).

1.3 Outline

This thesis consists of two main parts. The first part presents the background theory for vision-based global pose estimation. The second part contains three publications on the same topic.

1.3.1 Outline Part I: Background Theory

The background theory begins in chapter 2 with an introduction to the most common sensors and methods for aerial pose estimation and some details on useful geographic information to support the visual methods. Chapter 3 introduces the camera models employed for the aerial images used in this thesis. The basics for multiple view geometry are presented in chapter 4. Horizon detection and the concept of the Hough transform are described in chapter 5. Chapter 6 describes the fundamental methods utilized for vision-based pose estimation. The evaluation measures used are given in chapter 7 together with a description of how the ground truth pose was generated for the flights. Finally, the concluding remarks are given in chapter 8.

1.3.2 Outline Part II: Included Publications

Edited versions of three publications are included in Part II. The full details and abstract of these papers, together with statements of the contributions made by the author, are summarized below.

Paper A: Efficient 7D Aerial Pose Estimation

> B. Grelsson, M. Felsberg, and F. Isaksson. Efficient 7D Aerial Pose Estimation. *IEEE Workshop on Robot Vision*, 2013.

Abstract:
A method for online global pose estimation of aerial images by alignment with a georeferenced 3D model is presented. Motion stereo is used to reconstruct a dense local height patch from an image pair. The global pose is inferred from the 3D transform between the local height patch and the model. For efficiency, the sought 3D similarity transform is found by least-squares minimizations of three 2D subproblems. The method does not require any landmarks or reference points in the 3D model, but an approximate initialization of the global pose, in our case provided by onboard navigation sensors, is assumed. Real aerial images from helicopter and aircraft flights are used to evaluate the method. The results show that the accuracy of the position and orientation estimates is significantly improved compared to the initialization and our method is more robust than competing methods on similar datasets. The proposed matching error computed between the transformed patch and the map clearly indicates whether a reliable pose estimate has been obtained.

Contribution:
In this paper, a local height map of an urban area below the aircraft is computed using motion stereo. The global pose of the aircraft is inferred from the 3D similarity transform between the local height map and a geo-referenced 3D model of the area. The main novelty of the paper is a framework that enables the 3D similarity transform to be reliably and robustly estimated by solving three 2D subproblems. The author contributed to the design of the method, implemented the algorithms, performed the evaluation and the main part of the writing.

Paper B: Probabilistic Hough Voting for Attitude Estimation from Aerial Fisheye Images

> B. Grelsson and M. Felsberg. Probabilistic Hough Voting for Attitude Estimation from Aerial Fisheye Images. *18th Scandinavian Conference in Image Analysis, SCIA*, pages 478–488, 2013.

Abstract:
For navigation of unmanned aerial vehicles (UAVs), attitude estimation is essential. We present a method for attitude estimation (pitch and roll angle) from aerial fisheye images through horizon detection. The method is based on edge detection and a probabilistic Hough voting scheme. In a flight scenario, there is often some prior knowledge of the vehicle altitude and attitude. We exploit this prior to make the attitude estimation more robust by letting the edge pixel votes be weighted based on the probability distributions for the altitude and pitch and roll angles. The method does not require any sky/ground segmentation as most horizon detection methods do. Our method has been evaluated on aerial fisheye images from the internet. The horizon is robustly detected in all tested images. The deviation in the attitude estimate between our automated horizon detection and a manual detection is less than 1°.

Contribution:
This paper introduces one of only a few available methods using omnidirectional aerial images for absolute attitude estimation from horizon detection. The main novelty is the combination of (a) computing attitude votes from projection of edge pixels and their orientation on the unit sphere, and (b) weighting the votes based on the prior probability distributions for the altitude and pitch and roll angles, in order to obtain a robust and geometrically sound attitude estimate. The author contributed to the idea and design of the method, implemented the algorithms, conducted the evaluation and did the main part of the writing.

Paper C: Highly accurate attitude estimation via horizon detection

> B. Grelsson, M. Felsberg, and F. Isaksson. Highly accurate attitude estimation via horizon detection. *Submitted to Journal of Field Robotics*, 2014

Abstract:
Attitude (pitch and roll angle) estimation from visual information is necessary for

GPS free navigation of airborne vehicles. We propose a highly accurate method to estimate the attitude by horizon detection in fisheye images. A Canny edge detector and a probabilistic Hough voting scheme are used to compute an approximate attitude and the corresponding horizon line in the image. Horizon edge pixels are extracted in a band close to the approximate horizon line. The attitude estimates are refined through registration of the extracted edge pixels with the geometrical horizon from a digital elevation map (DEM), in our case the SRTM3 database. The proposed method has been evaluated using 1629 images from a flight trial with flight altitudes up to 600 m in an area with ground elevations ranging from sea level up to 500 m. Compared with the ground truth from a filtered IMU/GPS solution, the standard deviation for the pitch and roll angle errors are 0.04° and 0.05°, respectively, with mean errors smaller than 0.02°. The errors obtained are about one order of magnitude smaller than for any previous vision-based method for attitude estimation from horizon detection in aerial images. To achieve the high accuracy attitude estimates, the ray refraction in the earth atmosphere has been taken into account.

Contribution: This paper addresses the problem of attitude estimation from horizon detection in images. The paper presents the very first method where the attitude estimates from the horizon in omnidirectional images is refined through registration with the geometrically expected horizon from a digital elevation model. It is one of few methods where the ray refraction in the atmosphere is taken into account which contributes to the highly accurate pose estimates. The author planned and participated in the conduction of the field trials, contributed to the idea and design of the method, implemented the algorithms, performed the evaluation and the main part of the writing.

Other Publications

The following publications by the author are related to the included papers.

> B. Grelsson, M. Felsberg, and F. Isaksson. Global Pose Estimation of Aerial Images. *SSBA*, 2013. (Revised version of Paper A)

Chapter 2

Sensors and geographic information

Although this thesis is focused on *vision-based* methods for global pose estimation of airborne vehicles, it is essential to bear in mind that vision is not the most common onboard sensor used for pose estimation. The primary choice is most often inertial sensors and/or a GPS receiver. Hence, it is relevant to give a brief description of these standard sensors together with their capabilities and drawbacks. In addition, vision-based methods are often used in conjunction with these sensors in a filtering network. A conceivable sensor fusion network for vehicle pose estimation is shown in figure 2.1.

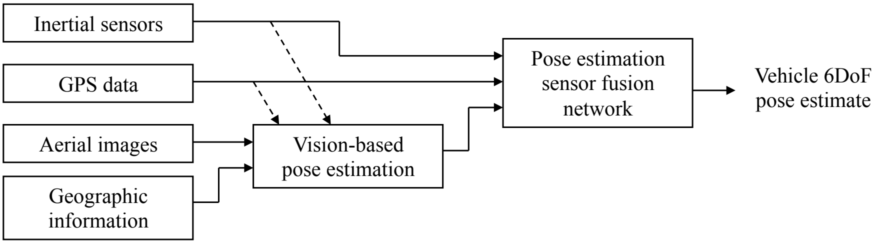
Inertial sensors
GPS data
Aerial images
Geographic information
Vision-based pose estimation
Pose estimation sensor fusion network
Vehicle 6DoF pose estimate

Figure 2.1: Block diagram for sensor fusion network for vehicle pose estimation. Information from inertial sensors and GPS may optionally be used as inputs to the vision-based pose estimation method.

2.1 Standard sensors for aerial pose estimation

A standard procedure for full six degree-of-freedom (6DoF) pose estimation (position and orientation) of an airborne vehicle is fusion of data from an inertial navigation system (INS) with data from a global navigation satellite system (GNSS) in a filtering network. To make the filtering more robust and accurate, it is also common to utilize magnetometers and altimeters as complementary sensors.

An INS contains as a minimum a computer, linear motion sensors (accelerometers) and rotation sensors (gyroscopes) to continuously compute the position, velocity and orientation. Since an INS measures accelerations and angular rates, it can detect changes in the position, velocity and orientation relative to its original state by integration of the measured values over time. This means that the estimated absolute position and orientation given by an INS are relative to the entities valid at a reference time. It also means that an INS suffers from integration drift, i.e. small errors in the acceleration and angular velocity measurements are integrated and accumulated to larger errors in velocity and orientation. Position estimates which comprise a double integration of the measured accelerations suffer even more from drift without support from other sensors. Many INSs also contain a magnetometer for heading measurements and a barometer for measuring the altitude.

An INS or an inertial measurement unit (IMU) is often categorized based on its accuracy over time, i.e. on its drift. High accuracy equipment with drift in the order of 1°/hour usually have a weight in the order of kilograms and may not be carried by small UAVs. Microelectromechanical system (MEMS) IMUs, which are considerably lighter, are more suited for these platforms. The penalty is a significantly larger drift, usually in the order of 10°/hour.

To alleviate the effect of drift from an INS, continuous absolute position estimates may be used in a filtering network. The most common way to provide this information is by utilizing a GNSS. A GNSS is a system of satellites providing autonomous geo-spatial positioning capability with global coverage. Today there are two operational global GNSSs, the widely spread american Global Positioning System (GPS) and the russian GLONASS. The global position of a receiver on earth is computed from simultaneous reception of microwave signals from several satellites orbitting the earth in known trajectories. The absolute position accuracy of a single measurement for an ordinary GPS receiver is in the order of a few meters. It is also common knowledge that the accuracy of a GPS degrades in the vicinity of tall buildings caused by blocked reception of the satellite signals (radio shadow) but also due to multipath effects, i.e. signals that have been reflected in other objects will also reach the receiver, not just direct signals from the satellite. GPS receivers can be made small and light weight and can generally be carried by small UAVs.

An air pressure meter is frequently used in aircraft to measure the altitude above sea level. The physical phenomenon exploited is that the air pressure drops almost exponentially with the altitude within the earth atmosphere. It is, however, a relative measurement since the sea level atmospheric pressure at a certain location varies over time with the temperature and movement of pressure systems in the atmosphere. Air pressure meters can be made small and light weight and can be carried by small UAVs.

2.2 Geographic information for vision support

The vision-based methods for global pose estimation presented in part II of this thesis use a geo-referenced 3D model and ground elevation data as support. This is the reason why these types of geographic information are described and illustrated below. Other examples of geographic information used to support vision-based methods are: databases with images captured at known 6DoF poses, appearance and detailed shape of buildings, readily observable synthetic landmarks along the planned flight trajectory, road network maps, and lake and coastline contours.

2.3 Digital Elevation Models

Digital elevation models (DEMs) contain information on the altitude of the terrain as a function of ground position. The altitude data is often given over an equidistant XY grid with a certain resolution, but it may also be represented with an irregular triangular network. DEMs may be generated from several types of airborne sensors like lasers, radars and cameras.

When using a DEM it is of course essential to know the accuracy of the altitudes given. But equally important is to know whether the altitudes represent the true surface (including houses and vegetation), the ground terrain or a mixture of these, see figure 2.2. This will depend on the sensor used as well as the postprocessing performed when generating the model. When matching altitudes computed from image data with a DEM, one error source may be the type of altitude the DEM actually represents, i.e. if it is a surface model, a terrain model or something in between.

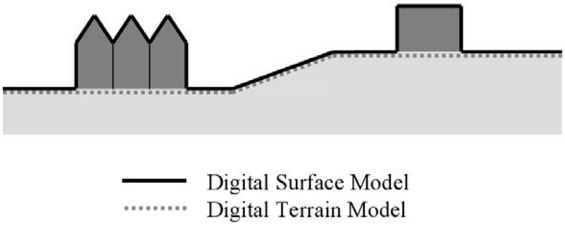

Figure 2.2: Surface and terrain model.

In paper C, access to elevation data over a region as large as ∼400x400 km was required. This data could be provided by the publicly available database SRTM3 [25]. SRTM stands for Shuttle Radar Topography Mission and the data was captured from the space shuttle Endeavour in the year 2000. The data acquisition mission took 11 days. The radar wavelength used was around 5 cm and the cell grid resolution is ∼90x90 m. Although a large part of the microwaves at that wavelength was reflected on the surface of the objects within a cell, the altitudes given in the database will represent some sort of altitude average over the area covered within a cell. The absolute altitude accuracy is claimed to be around 8-10 m. Elevation data from the SRTM3 database is shown in figure 2.3.

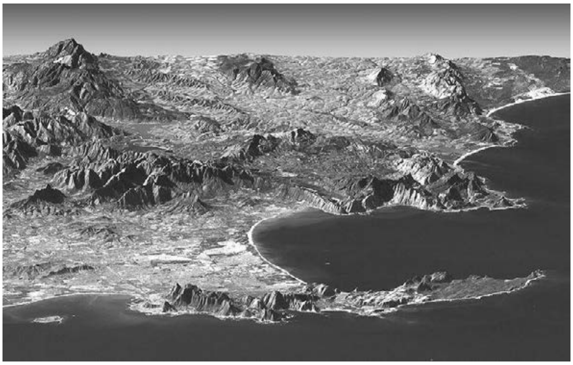

Figure 2.3: Elevation data from SRTM3 database from NASA website. The elevation data is textured with LANDSAT imagery.

The purpose of SRTM was to generate an elevation model with global coverage. A large distance between the sensor and earth was used for coverage, at the same time reducing the cell grid resolution. Generating a large scale elevation model at high resolution by flying at low altitude (<1km) with a small footprint is very time consuming. One such example is laser scanning the entire of Sweden to generate NNH (Ny Nationell Höjdmodell - new national elevation model) with a 2 m cell grid resolution. The ongoing project is planned to take around 5 years. Elevation data in NNH will represent the ground terrain, which is possible to filter out since the laser will obtain reflections from both the top of the vegetation and the underlying ground.

The purpose of presenting the acquisition times for the DEMs is to reflect that generating a highly accurate, dense grid DEM over a large area is a major project and although the available DEMs may not represent the ideal altitude information for a certain vision-based application, they may still be the only practical choice.

2.4 Vision-based 3D models

Elevation models and 3D models may also be generated from aerial imagery using a process called structure from motion. Images are captured over the desired area from an airborne vehicle which may be a satellite, an aircraft or a UAV, depending on the model resolution and coverage desired.

In paper A, a 3D model generated with Saab's Rapid 3D Mapping^{TM} process [32] was used. The georeferenced model was created from images captured from an aircraft scanning the city of Linköping at 600 m altitude. The 3D data is represented with a triangular mesh onto which texture from the images is overlaid to give a photorealistic impression. This feature also makes it possible to render images from a virtual camera at any position in the model, a method often used when trying to register a new image with the 3D model for global pose estimation. Figure 2.4 shows an example of an urban area in Linköping from a Rapid 3D Mapping model.

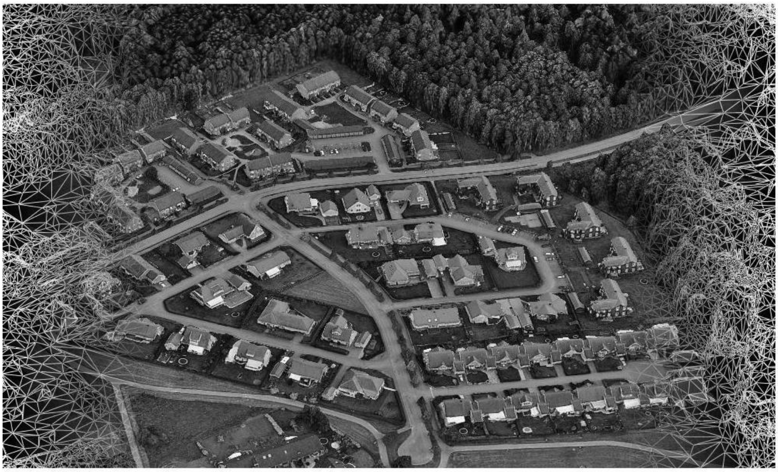

Figure 2.4: Rapid 3D Mapping model with triangular mesh and overlaid texture.

2.5 Ray refraction in the atmosphere

It is common knowledge that the air pressure drops with the altitude above the ground, i.e. the air gets thinner at higher altitudes. It may be less known that this fact affects the exact location of the perceived horizon.

It is not only the air pressure that varies with the altitude when the air gets thinner. The refractive index n exhibits a similar behavior. The refractive index is higher close to the ground and drops with the altitude and equals unity in vacuum. This means that the light rays from an aircraft to the perceived horizon will not be straight lines but instead they will be slightly bent. This is illustrated in figure 2.5.

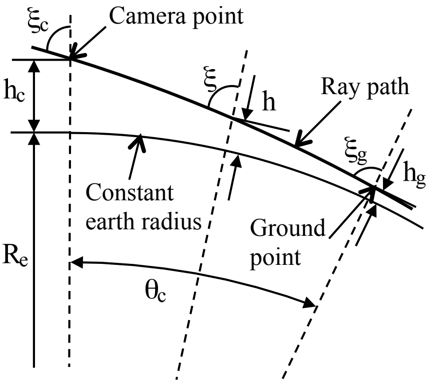

Figure 2.5: Refracted ray path in the atmosphere.

To quantitatively model the ray bending effect we first assume the earth to be a sphere with radius R_e. Overlaid on this sphere there is ground topography. We then use a spherically stratified model with thin radial layers of thickness Δr and with constant refractive index. Using Snell's law, it can be shown [16] that a ray propagating through a spherically stratified model will follow a path which obeys the equation

$$nr\sin\xi = n_\mathrm{c}r_\mathrm{c}\sin\xi_\mathrm{c} = n_\mathrm{g}r_\mathrm{g}\sin\xi_\mathrm{g} = k = \text{constant} \tag{2.1}$$

where $r = R_\mathrm{e} + h$ is the radius and ξ is the incidence angle in the layer. The subscripts c and g denote the camera and the ground, respectively. The refractive index n as a function of the altitude h can be modelled, as a first order approximation, as

$$n(h) = 1 + A\exp(-h/B) \tag{2.2}$$

where A and B are not truly constants but vary slightly with the current air pressure and temperature.

For the ideal horizon at sea level we have $h_\mathrm{g} = 0$ and $\xi_\mathrm{g} = \pi/2$. Using (2.1) and (2.2), the incidence angle ξ_c for the ideal horizon on a camera at altitude h_c can be determined. In paper C, equations are given for the angle θ_c from which the distance to the ideal horizon can be computed. Further, paper C also details how to compute the incidence angle to the perceived horizon when there are objects at altitudes above the ray path from the camera to the ideal horizon at sea level.

Chapter 3

Camera Models

Vision-based pose estimation relies on the fact that accurate mathematical relationships can be established between 3D points in a world coordinate frame and their corresponding image coordinates. As the name suggests, *camera models* are used to provide a mathematical model how light rays from an object are propagated through the camera lens to the sensor chip, where the rays create an image. In paper A, cameras with a *normal* lens with fixed focal length was used. *Normal* in this context refers to the fact that the image is close to what we humans normally see with our own eyes. For this type of lens, a simple pinhole camera model with minor lens distortion corrections is often adequate. For a *fisheye* lens, which was used in papers B and C, the true lens design is far more complex and this is also reflected in the lens model which mathematically is more involved than for a normal lens.

3.1 Pinhole camera model

The very first camera used to acquire an image, a *camera obscura* [33], used the principle of a pinhole camera. Light rays from the object passed through an infinitesimal hole and created an image on the wall inside a dark box. The geometry is illustrated in figure 3.1 where, for mathematical convenience, the image plane has been placed in front of the pinhole and not behind it inside the box.

Consider a world 3D point $\mathbf{X} = [x \;\; y \;\; z]^T$. If the distance from the pinhole to the image plane is denoted d, the image coordinates $\mathbf{u}$ for the point will be

$$\mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix} = \frac{d}{z} \begin{pmatrix} x \\ y \end{pmatrix} \tag{3.1}$$

It is often convenient to consider an image plane at unit distance from the pinhole, the so called normalized image plane. The normalized image coordinates are given by

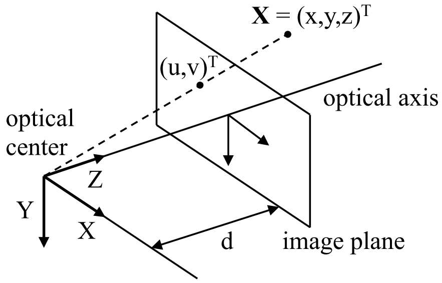

Figure 3.1: Pinhole camera model.

$$\mathbf{u_n} = \begin{pmatrix} u_\text{n} \\ v_\text{n} \\ 1 \end{pmatrix} = \begin{pmatrix} x/z \\ y/z \\ z/z \end{pmatrix} \tag{3.2}$$

In the real world, the pinhole is replaced with a thin lens (or rather a system of lenses) to allow for a larger aperture, letting more light through, and focus the light at a focal plane. Replacing the distance d with the focal length f for the lens, the pinhole camera model reads

$$\lambda \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} f & \gamma & u_0 \\ 0 & \alpha f & v_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{KX} \tag{3.3}$$

Since the sensor elements may not be exactly quadratic, an aspect ratio α has been introduced. The sensor may not be perfectly aligned with the lens allowing for a skew angle γ. For well manufactured lenses, α is very close to 1 and γ is negligible. The origin in the image plane is not along the optical axis but has an offset (u_0, v_0). λ is a scaling parameter. The linear mapping $\mathbf{K}$ is called the *intrinsic* camera matrix.

In general, the camera coordinate system is not aligned with the world coordinate system. Their interrelationship is described by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. These two parameters are called *extrinsic* camera parameters. If we define the camera matrix

$$\mathbf{C} = \mathbf{K}\left[\mathbf{R} \mid -\mathbf{Rt}\right] \tag{3.4}$$

we can formulate a linear mapping of a world point to its image coordinates as

$$\lambda \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \mathbf{C} \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix} \tag{3.5}$$

3.2 Lens distortion

The camera model in the previous section assumes a rectilinear projection of points, i.e. straight lines in the world will be projected as straight lines in the image plane. For a real-world lens this is not perfectly true although a very good approximation. Even though multiple-lens systems are used by the lens manufacturers, there still remains some imperfections called optical aberrations. The most common ones are radial and tangential distortion, chromatic and spherical aberration and astigmatism.

The physical characteristics of the lens system introduce a phenomenon called radial distortion. Typically a square object will be imaged either with a *barrel* distortion or a *pincushion* distortion as illustrated in figure 3.2. The image will be more distorted the further away from the center of the image. Tangential distortion in the image is created when the optical axis of the lens system is not perfectly aligned with the normal vector of the image sensor plane.

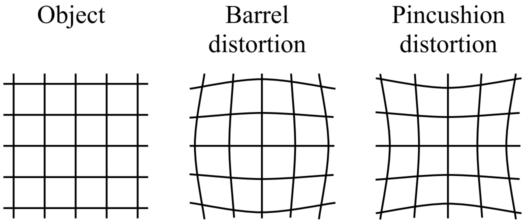

Figure 3.2: Radial distortion, barrel and pincushion.

The camera models used in papers A and C took radial and tangential lens distortions into account. To mathematically compensate for radial and tangential lens distortions we first define a set of undistorted coordinates for each point in the normalized image plane,

$$u_{\mathrm{u}} = u_{\mathrm{n}} \tag{3.6a}$$

$$v_{\mathrm{u}} = v_{\mathrm{n}} \tag{3.6b}$$

where the subscript u means *undistorted*. We then define the radial distance r as the distance from the origin in the normalized image plane. The total distortion in the x and y directions for each point in the normalized image plane is given by

$$r^2 = u_{\mathrm{u}}^2 + v_{\mathrm{u}}^2 \tag{3.7a}$$

$$du = u_{\mathrm{u}} \sum_i k_i r^i + 2t_1 u_{\mathrm{u}} v_{\mathrm{u}} + t_2(r^2 + 2u_{\mathrm{u}}^2) \tag{3.7b}$$

$$dv = v_{\mathrm{u}} \sum_i k_i r^i + 2t_2 u_{\mathrm{u}} v_{\mathrm{u}} + t_1(r^2 + 2v_{\mathrm{u}}^2) \tag{3.7c}$$

where the first term (coefficients k_i) is the radial distortion and the latter terms (coefficients t_i) comprise the tangential distortion. We denote the set of lens distortion parameters with D. In paper A, a polynomial of degree four was used for the radial distortion. For the fisheye lens in paper C, a polynomial of degree eight (even terms) was used.

The distorted coordinates in the normalized image plane are given by

$$u_{\mathrm{d}} = u_{\mathrm{u}} + du \tag{3.8a}$$

$$v_{\mathrm{d}} = v_{\mathrm{u}} + dv \tag{3.8b}$$

To obtain the final image coordinates for a pinhole camera with radial and tangential lens distortion, a mapping with the intrinsic camera matrix $\mathbf{K}$ is applied to the distorted coordinates.

Equations 3.7 - 3.8 give explicit expressions how to compute the forward lens distortion, i.e. going from undistorted to distorted coordinates. To compute the backward lens distortion, i.e. going from distorted to undistored coordinates, iterative methods are normally used to solve a nonlinear equation system.

3.3 Omnidirectional cameras

A perspective or normal lens, as described in the previous section, is aimed at imaging straight objects in the world as straight lines in the image. There is a completely different class of cameras called omnidirectional cameras. As the name suggests, the aim is now to obtain an omnidirectional or 360° view of the surroundings captured in one image. In this category there are two types of cameras; catadioptric cameras and fisheye cameras. Example images with these types of cameras are shown in figure 3.3. As expected, omnidirectional images are heavily distorted radially when projected on a plane.

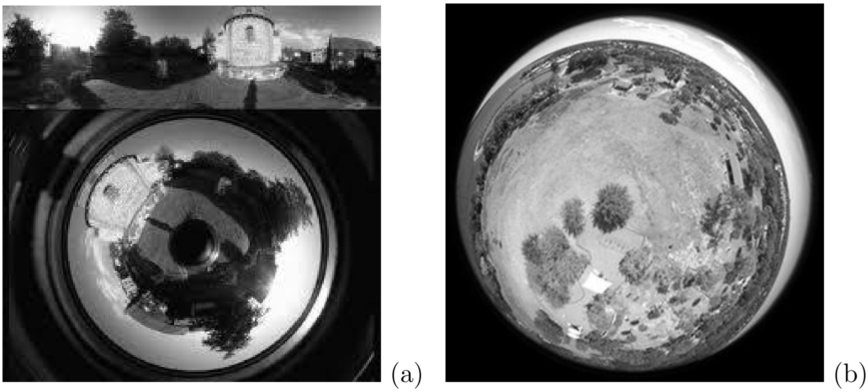

Figure 3.3: (a) Catadioptric image and panorama projection. (b) Fisheye image.

A catadioptric camera uses a reflective mirror with hyperbolic shape for image formation which results in a very characteristic image with a black center circle due to blocking from the sensor chip. A fisheye camera uses a system of lenses to achieve the aim of refracting light rays from roughly a hemisphere to a plane, see figure 3.4. Compared to a catadioptric camera, a fisheye camera is more complex in its design but it is smaller in size and the black center spot is avoided. A fisheye lens often suffers from noticeable chromatic abberation. A fisheye lens with a field of view larger than 180° creates very typical images with a *fisheye circle*, a border line on the image plane outside of which no light rays will reach the sensor due to geometrical constraints.

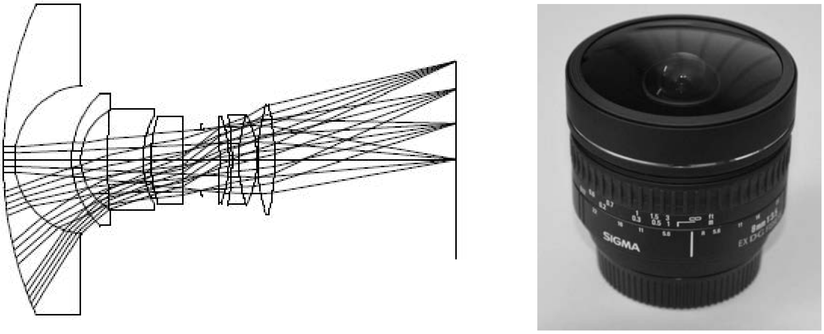

Figure 3.4: Fisheye lens design with refraction of light rays (Laiken, 1995). Fisheye lens used in flight trial for paper C.

3.4 Fisheye camera model

The fisheye camera model used in papers B and C is based on the aim of the fisheye lens design - to image a hemisphere of world points onto a plane. First, a 3D world point $\mathbf{X}$ is projected onto the unit sphere, placed at the camera location, as point $\mathbf{x}_s$, see figure 3.5. The point on the unit sphere is then projected onto the normalized image plane by a pinhole camera model with its optical center the distance L from the center of the unit sphere and focal distance 1 to the plane. Next, radial and tangential lens distortions are applied. The final projection is a generalized camera projection $\mathbf{K}$ given by the intrinsic camera parameters.

3.5 Pinhole camera calibration

The accuracy of the pose estimation methods we are interested in will of course be reliant on the accuracy of the camera calibration. For calibration of a perspective lens with a pinhole camera model, the method by Zhang [34] is often used. The calibration object is a planar grid or checkerboard pattern with known dimensions. Images of the calibration object are captured in different orientations. From the linear mapping (homography) between the grid points in the object plane and the image plane, constraints can be established on the camera intrinsic parameters. If

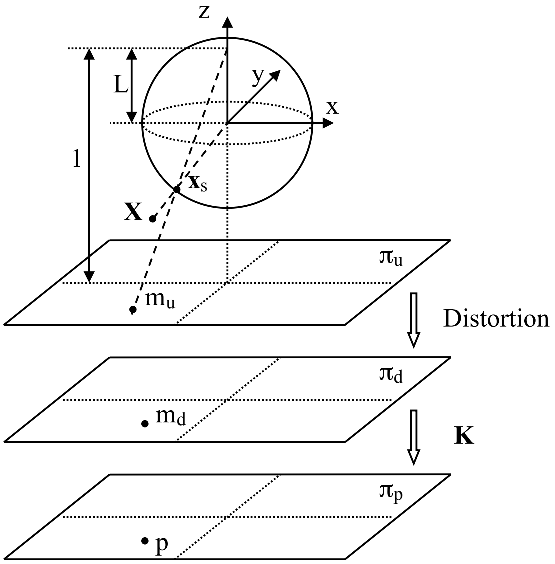

Figure 3.5: Fisheye camera model.

at least three independent orientations are used, all intrinsic and extrinsic camera parameters for a pinhole camera can be solved in closed form.

This calibration can be refined, also taking the lens distortion parameters into account, by minimizing the total reprojection error for all corner points,

$$(\mathbf{K}_{\text{est}}, D_{\text{est}}) = \arg\min_{K,D} \sum_{i=1}^{n} \sum_{j=1}^{m} \|\mathbf{u}_{ij} - \tilde{\mathbf{u}}(\mathbf{K}, \mathbf{R}_i, \mathbf{t}_i, \mathbf{X}_j, D)\|^2 \tag{3.9}$$

The summation is made over the camera positions i with the corresponding rotation $\mathbf{R}_i$, translation $\mathbf{t}_i$, the world corner points $\mathbf{X}_j$, the camera matrix $\mathbf{K}$ and the lens distortion parameters D. The symbol $\tilde{\mathbf{u}}$ denotes projection of a world point onto the image plane and $\mathbf{u}_{ij}$ are the true image points.

The accuracy of the calibration will be dependent on the subpixel accuracy for the detector when extracting the corner points on the calibration pattern. The method is widespread since it is sufficiently accurate for most applications and because the calibration pattern can be readily obtained from printers.

3.6 Fisheye camera calibration

Methods are also available for calibrating omnidirectional cameras with a checkerboard pattern [28], [24]. The method in [24] attempts to fit image data of a checkerboard pattern to the same fisheye camera model as presented in section 3.4. The method was used as a first stage for the camera calibration in paper C. For the fisheye camera model, the mapping from the world checkerboard plane to the image plane is not linear and no closed-form solution can be obtained for the calibration parameters. In [24] they use reasonable assumptions on some calibration parameters as an initialization and then minimize a similar error function as in (3.9) using the Levenberg-Marquardt algorithm [19].

In paper C, the calibration method in [24] was used to obtain an initial calibration subsequently refined using registration with true world 3D points. From the onboard navigation sensors an accurate ground truth for the vehicle 6DoF pose is available. Given the world 3D position for the camera, a geometrical horizon projected onto the unit sphere can be computed from DEM data. If horizon pixels can be extracted from the images, all information is available to compute a refined camera calibration using (3.9). The calibration method proposed in paper C solved a dual formulation, i.e. it minimized the distances between the corresponding points on the unit sphere and not on the image plane.

Chapter 4

Multiple view geometry

When a camera mounted on an airborne vehicle captures images of the ground at high frame rates, there will generally be a substantial image content overlap between successive images. Geometrically, the combined image content from an image pair can be utilized analogously to how our human vision system uses stereo images. In the same way that we humans can determine distances and directions to objects within our view, the same information can be determined from two images if we know the stereo baseline (distance between the eyes) and how the cameras (eyes) are oriented. The principle for this two-view geometry, or *epipolar geometry*, is one of the keystones in computer vision and the fundament for vision-based reconstruction of 3D structures. An example of two aerial images, used in paper A, with some image point correspondences is shown in figure 4.1.

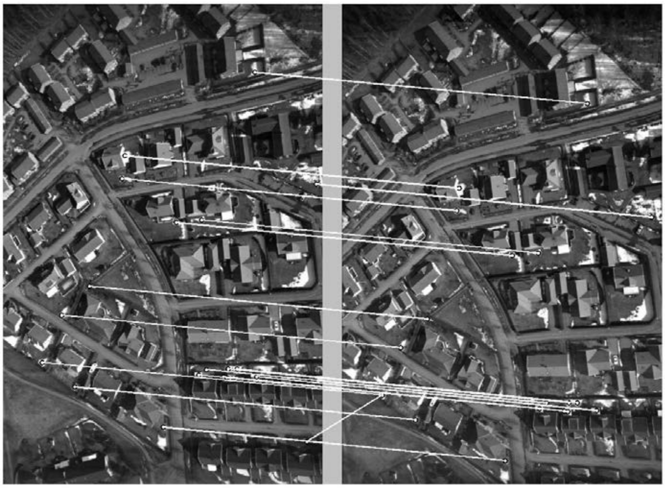

Figure 4.1: Image point correspondences and one feature matching outlier.

4.1 Epipolar geometry

The classical way to explain the concept of epipolar geometry is to consider the geometry in figure 4.2. Two pinhole cameras, located at positions $\mathbf{O}_1$ and $\mathbf{O}_2$, are imaging the same world point $\mathbf{X}$. The two cameras may be the same physical camera that has been moved or two different cameras, but the camera center locations need to be distinct. The projection of the world point $\mathbf{X}$ on the two image planes will be at $\mathbf{u}_1$ and $\mathbf{u}_2$ respectively. But $\mathbf{u}_1$ will also be the image point for all points on the 3D line passing through $\mathbf{O}_1$ and $\mathbf{X}$, e.g. the world points $\mathbf{X}'$ and $\mathbf{X}''$. In the second camera, this 3D line will be imaged as the line $\mathbf{l}_2$ called an *epipolar line.* Repeating this process for other world points, it can be shown that all epipolar lines in the second image will intersect at a point $\mathbf{e}_2$ called the *epipole.* The epipole $\mathbf{e}_2$ is also the image point in the second image of the camera center $\mathbf{O}_1$.

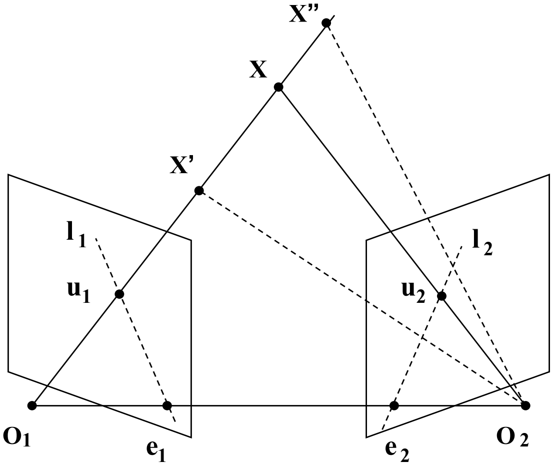

Figure 4.2: Epipolar geometry.

The constraint that an image point in the first image must lie on an epipolar line in the second image is known as the *epipolar constraint* and can mathematically be expressed as

$$\mathbf{u}_1^T \mathbf{l}_2 = \mathbf{u}_1^T \mathbf{F} \mathbf{u}_2 = \mathbf{0} \ . \tag{4.1}$$

$\mathbf{F}$ is called the fundamental matrix and is a 3x4 matrix with seven degrees of freedom. A thorough mathematical derivation of the epipolar constraint and how the matrix $\mathbf{F}$ is related to the camera matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ can be found in [17].

If two images have been captured in distinct locations with a calibrated camera, the expressions for the epipolar constraint can be simplified further. We now denote the normalized coordinates for a point in the first image as $\mathbf{u} = [u \ \ v \ \ 1]^T$ and in the second image as $\tilde{\mathbf{u}} = [\tilde{u} \ \ \tilde{v} \ \ 1]^T$. For corresponding image points with a calibrated camera, the epipolar constraint can be expressed as

$$\mathbf{u}^T \mathbf{E} \tilde{\mathbf{u}} = 0 \ . \tag{4.2}$$

The matrix $\mathbf{E}$ is called the essential matrix and can be decomposed as

$$\mathbf{E} = [\mathbf{t}_{12}]_{\mathrm{x}}\mathbf{R}_{12}^T \tag{4.3}$$

where $\mathbf{R}_{12}$ and $\mathbf{t}_{12} = [t_x \;\; t_y \;\; t_z]^T$ are the relative rotation and the translation for the camera between images 1 and 2, and $[\cdot]_{\mathrm{x}}$ denotes the cross-product operator meaning that

$$[\mathbf{t}_{12}]_{\mathrm{x}} = \begin{pmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{pmatrix}. \tag{4.4}$$

The essential matrix has five degrees of freedom. This means that the matrix $\mathbf{E}$ can be computed from a set of five corresponding image points. Since $\mathbf{R}_{12}$ and $\mathbf{t}_{12}$ jointly have six degrees of freedom, it also means that from corresponding image points it is feasible to compute the relative rotation and the direction of the translation vector but not its magnitude. This is known as the *scale ambiguity*, i.e. the size of objects in the two images cannot be deduced from image information alone. An implementation of a five-point-solver for $\mathbf{E}$ can be found in [26] which also presents how to decompose the essential matrix to a rotation and a translation.

4.2 Local pose estimation

The pose estimation method in paper A is based on the computation of a local height map from two images and registration of the height map with a 3D model of the area. To compute the local height map, the relative rotation and translation between the two images need to be determined.

From section 4.1 we know that the essential matrix $\mathbf{E}$ can be computed from at least five image correspondences. In the problem formulation in paper A, an airborne monocular camera was supported by an IMU providing very accurate estimates of the relative rotation between images. If the rotation matrix $\mathbf{R}_{12}$ is known, the epipolar constraint can be expanded using (4.3 - 4.4) to yield a linear equation system for the translation $\mathbf{t}_{12} = [t_x \;\; t_y \;\; t_z]^T$ given by

$$\begin{bmatrix} (r_{12}u + r_{22}v + r_{32}) - \tilde{v}(r_{13}u + r_{23}v + r_{33}) \\ \tilde{u}(r_{13}u + r_{23}v + r_{33}) - (r_{11}u + r_{21}v + r_{31}) \\ \tilde{v}(r_{11}u + r_{21}v + r_{31}) - \tilde{u}(r_{12}u + r_{22}v + r_{32}) \end{bmatrix}^T \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} = 0 \tag{4.5}$$

where r_{ij} are the components of the rotation matrix. This linear equation system can be solved in a least-squares sense using a singular value decomposition (SVD) from two or more image correspondences. Some questions then raised are: How can these image correspondences be obtained? How many of these correspondences, and which ones, shall be used for the pose estimate, in this case the estimate of $\mathbf{t}_{12}$? These questions will be answered in the next two sections.

4.3 Feature points

A feature point is an image point contained in a small image region that possesses some characteristics that deviate from its surroundings. Some common methods for feature point detection are SIFT [20], SURF [3], FAST [27] and Shi-Tomasi [30]. To find corresponding image points between two images either feature matching, feature tracking or a combination of the two may be used.

For feature matching, a descriptor is computed for each feature point. The descriptors may e.g. be based on image intensities, color, gradient direction, orientation histograms, scale, etc. The next step is to match feature points in the two images. A distance measure in the descriptor space is defined and feature points having a small distance (below a given threshold) are potentially a good match.

In feature tracking, a feature point is detected in the first image. This point is tracked, e.g. with a Lucas-Kanade tracker [21], cross correlation or block matching, to find the corresponding location in the next image. In paper A, a combination of a FAST detector and a KLT-tracker was employed to find potential image correspondences.

In general, the feature matching or feature tracking process will generate a large number of image correspondences, often several hundreds depending on the thresholds set. Inevitably, among the potential image correspondences, there will be mismatches. If these mismatches are included in the solution of $\mathbf{t}_{12}$, they will induce an erroneous estimate. This is where the concept of RANSAC [9], RANdom SAmple Consensus, is widely used to obtain a robust estimation.

4.4 RANSAC

The main idea behind RANSAC is readily explained estimating a best linear fit of an almost linear point cloud in 2D. Consider a point set originating from an experiment in physics where the expected relationship between two parameters is linear. Due to measurement errors there is a spread around the expected line and, in addition, there are also some truly bad measurements, so called outliers. Using RANSAC to obtain a robust estimate of the line, two random points in the point cloud are chosen. A line is drawn between these two points. An accepted error distance from this line is defined and all points within the error margin, the inliers, constitute the consensus set. This process, randomly picking two new points each time, is repeated for a certain number of times, and the line hypothesis with the largest consensus set is a robust estimate of the line. A line estimate based on all points from the largest consensus set may be used to further refine the line estimate.

Applying RANSAC for the estimate of $\mathbf{t}_{12}$ in (4.5), two randomly selected image correspondences from the full matching set were chosen to compute an estimate $\mathbf{t}_{\text{est}}$ and the corresponding estimate $\mathbf{E}_{\text{est}}$ of the essential matrix. To compute the consensus set, a common choice for the error measure is the distance in the image plane between an image point and the epipolar line computed from $\mathbf{E}_{\text{est}}$ and the corresponding image point. After repeating this process for a desired number of times, the points in the largest consensus set were used to compute the final estimate $\mathbf{t}_{\text{est}}$.

4.5 Structure from Motion

Structure from motion (SfM) was used as a tool in the overall method for global pose estimation in paper A. A brief description of the concept of SfM and 3D reconstruction is given below.

From epipolar geometry, we know that the relative rotation and translation between two images for a calibrated camera can be computed from image correspondences. Contrary, if we do know the rotation and the absolute translation for the camera between images, it is possible to compute the 3D coordinate for the world point corresponding to the image points. In figure 4.2, consider the line passing through the points $\mathbf{O}_1$ and $\mathbf{u}_1$ and a second line passing through the points $\mathbf{O}_2$ and $\mathbf{u}_2$. The intersection between these two lines will yield the 3D coordinate for the world point $\mathbf{X}$. This method to determine the position of the world point from two image points is called *triangulation*. In reality, these two lines will rarely intersect due to e.g. image noise, nonperfect feature tracking/matching and often the 3D point is taken as the point where the distance between the two lines is the shortest.

When using stereo vision for 3D reconstruction of a scene, the reconstruction can either be made sparse or dense. A sparse reconstruction means that a limited set of corresponding image points (feature points) are used to compute 3D points. For a dense reconstruction, all image points jointly seen in the two images are used to compute the 3D points.

4.6 Sparse 3D reconstruction

For a sparse 3D reconstruction, feature points are first detected in two images. Image correspondences are established with feature tracking or feature matching and the corresponding 3D points are computed from triangulation given the assumption on the relative rotation and translation for the two cameras. This will yield a sparse set of 3D points obtained from two images.

If a third image is available, image correspondences are established between the third image and the first two images. The 3D points computed from the first two images are projected into the third image and minimizing the reprojection error for these points, keeping the 3D points fixed, the camera pose for image three can be estimated. This process is called Perspective-N-Point (PNP) pose estimation.

Triangulation is then performed to generate new 3D points for image correspondences not previously used. The camera poses and the position of all 3D points can then be optimized by minimizing the reprojection errors, i.e. the distance between the true image points and the projection of the 3D points on the image plane. This overall 3D point and camera pose optimization is called *bundle adjustment*.

This process adding new images, performing PNP, triangulation and bundle adjustment may be repeated to obtain a 3D reconstruction based on a large number of images.

4.7 Dense 3D reconstruction

In paper A, a dense 3D reconstruction was computed using a phase-based stereo algorithm, based on [11], and further developed at SAAB. The *disparities* obtained are converted to distances based on the estimated stereo baseline, the distance between the two cameras. The 3D reconstruction is then ortho-rectified based on the assumed global pose for camera 1 to generate a local height map. An example height map from an urban area is shown in figure 4.3.

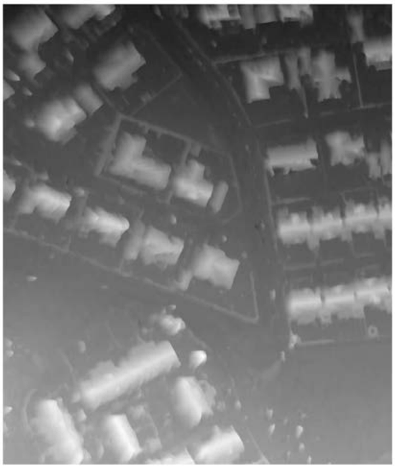

Figure 4.3: Dense 3D reconstruction of an urban area. The intensity represents height above the ground.

Chapter 5

Horizon detection

As mentioned in the introduction, the horizon line is often used by pilots as an attitude guide during takeoff and landing. In the same manner, horizon detection is an intuitive means for us humans to determine the absolute attitude (pitch and roll angles) whenever up in the air. This insight has led to a large number of vision-based methods for attitude estimation via horizon detection in aerial images.

The image shape of the horizon will depend largely on the type of camera used on the aircraft. For front-mounted perspective cameras, the horizon (sky/ground border) will generate a contiguous rather straight line across the image. For omnidirectional cameras, the horizon will ideally form an ellipse on the image plane. This is illustrated in figure 5.1. Due to blocking by the fisheye circle, the horizon does not generate a complete ellipse on the fisheye image plane.

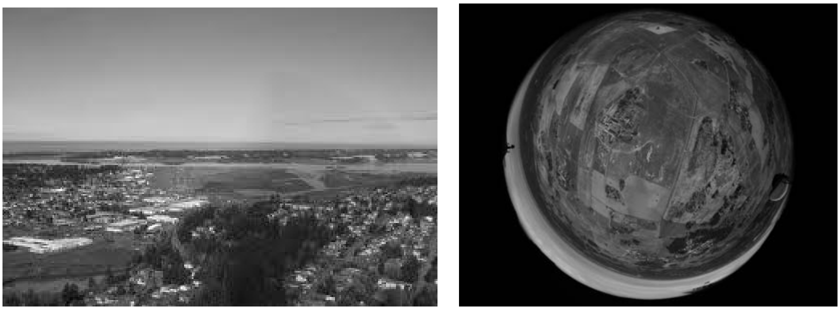

Figure 5.1: Horizon seen in perspective image and fisheye image.

Irrespective of the camera type used, there are two main strategies for horizon detection in images. For one group of methods, sky/ground *segmentation* of the image is the key component. The segmentation may e.g. be based on pixel color and/or texture content. Horizon detection methods based on sky/ground segmentation are found in [31] and [23] for perspective images and in [6] for omnidirectional images. Once the horizon line has been determined from the segmentation process, its location in the image is converted to attitude angles.

The second group of methods relies on the fact that the horizon should generate a distinct edge in the image. A commonly used method to detect edges is the Canny detector [5]. Horizon detection methods based on edge detection can be found in [2], [8] and [7] for perspective images. Papers B and C are the first papers proposing methods for horizon detection in omnidirectional images based on edge detection. One difficulty for edge based methods is to select which edge pixels originate from the horizon and which edge pixels are generated by the remainder image scene content.

A common approach to detect lines or ellipses in images is to use the concept of Hough voting, a democratic process with a majority decision taken by all edge pixels. An advantage of Hough voting is that the sought shape can be detected although the complete shape is not present in the image. Hough voting was used in papers B and C, and the main principle for the process, also known as the Hough transform, is given below.

The reason for choosing omnidirectional images for horizon detection and attitude estimation in papers B and C is primarily that a substantially larger portion of the horizon can be seen in the image compared with a perspective image. It is rather obvious that a more accurate attitude estimate can be obtained when having information available on half of the full horizon compared to seeing only a tenth of it.

5.1 Hough circle transform

The Hough transform is a method for detecting a curve of a specified shape in an image. It exploits the duality between points on a curve and the parameters for that curve. The original work on the Hough transform was presented in [18] and was restricted to binary edge images to detect the shape of interest. Straight lines and circles are the most common shapes to be detected with the Hough transform. The generalized Hough transform [1] enables instances of any arbitrary shape in an image to be detected.

Since the horizon detection methods in papers B and C are based on circle detection on the unit sphere, we explain the concept behind the Hough transform to detect an instance of a circle in a 2D image.

The well known equation for a circle is

$$(x - x_0)^2 + (y - y_0)^2 = r^2. \tag{5.1}$$

where the three parameters, the center point (x_0, y_0) and the radius r fully describe the circle in 2D.

The example image, in which we want to detect a circle using the Hough transform, is shown in figure 5.2 a). *Detect* in this context means that we want to determine the three parameters describing the circle. Symbolically, the image contains a dark earth surrounded by a bright sky. Running an edge detector on the image will yield a binary edge map as in figure 5.2 b). For this simple image, thresholding the magnitude of the gradient would yield the same edge map.

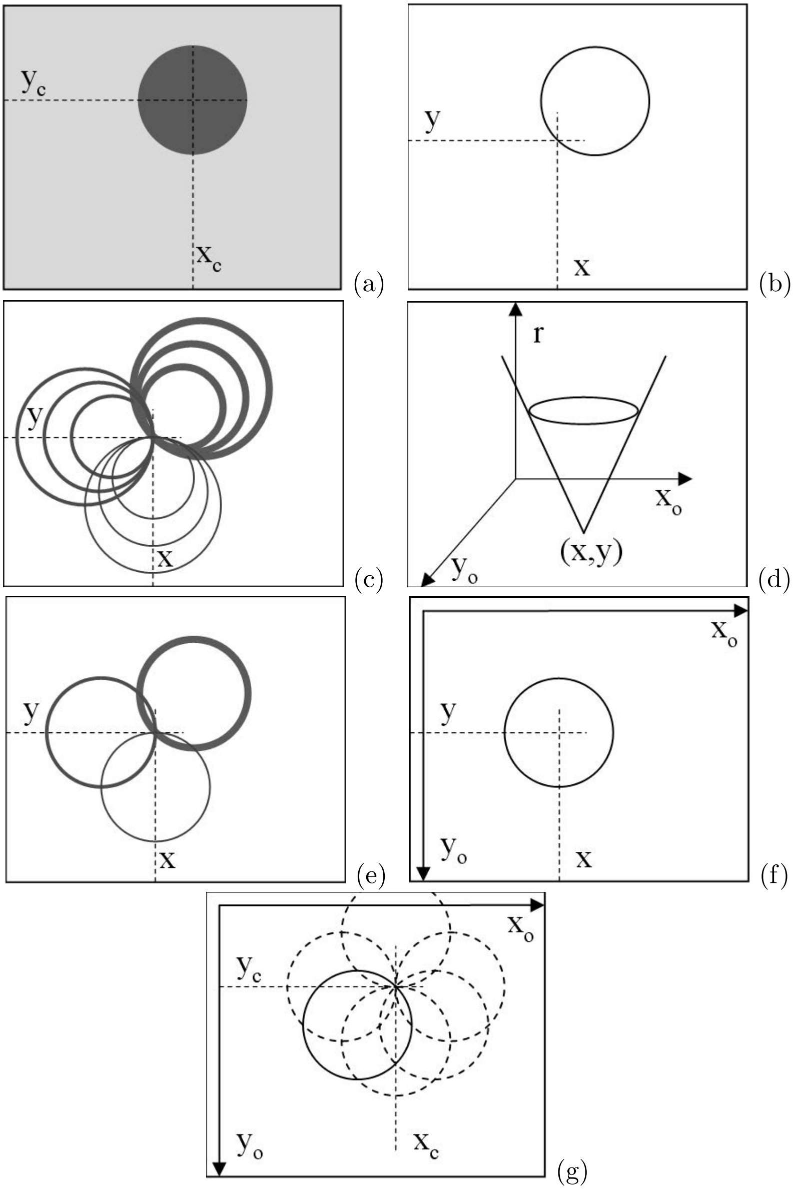

Figure 5.2: Hough transform. (a) Original image. (b) Edge map. (c) Possible circles; center point and radius unknown. (d) Same as (c) but in parameter space. (e) Possible circles; center point unknown, radius known. (f) Hough voting for one point if radius is known. (g) Hough voting for several points if radius is known.

We now consider the specific pixel (x, y) in figure 5.2 b) on the edge map. Knowing that the circle passes through (x, y) restricts the possible circles in the image to two degrees of freedom. Some examples of possible circles passing through (x, y) are shown in figure 5.2 c). The circle center point could lie along any line emanating from the point (x, y). Here, three example sets are shown with thick, medium and thin circle lines. For each direction, the circle radius could have any size, here illustrated with a small, a medium and a large size circle. In parameter space, the possible circles passing through (x, y) create a cone with its tip in (x, y) as shown in figure 5.2 d).

Now, assume that we are searching for circles with medium size radius r_{m}. This enforces another constraint on the possible circles and only one degree of freedom remains. A set of possible circles remaining in image space is shown in figure 5.2 e). In parameter space, the full set of possible circle center points create a circle with radius r_{m} around the point (x, y) as shown in figure 5.2 f). In the Hough voting, the edge point (x, y) would give votes for all center points along this circle. We now let all points along the edge map to vote for the circle center point. Each edge point will vote for center points along a circle as shown in figure 5.2 g). Aggregating the votes from all edge points, it is clear that the true circle center $(x_{\mathrm{c}}, y_{\mathrm{c}})$ will receive votes from all edge pixels and obtain the highest score in the accumulator array (voting grid) in parameter space. A search for the global maximum in the accumulator array enables the true parameters for the circle in the image to be determined.

It is also common to use the gradient direction as a constraint in the Hough transform. If we compute the gradient direction in the point (x, y) and enforce that the circle center point must lie along the gradient direction from the point (x, y), it is obvious that only the thick circle would remain in figure 5.2 e) and that only the true center point $(x_{\mathrm{c}}, y_{\mathrm{c}})$ would receive a vote in the parameter space in figure 5.2 g).

5.2 Hough transform - ellipse detection

In papers B and C, the objective was to detect the horizon in fisheye images to deduce the camera pitch and roll angles. For a fisheye camera and assuming a smooth earth, the horizon line will generate a circle when projected onto the unit sphere. On the image plane, it will form an ellipse, see figure 5.3.

The available information to estimate the attitude angles is an edge map, a camera calibration and we also assume the flight altitude h to be known, e.g. from a pressure meter onboard the airborne vehicle. Somehow we want to exploit the location of all edge pixels and their gradient direction to vote for a camera attitude. Which is the most tractable parameter space for the Hough voting? We decided to perform the voting on the unit sphere and not on the image plane. Why? Here are some qualitative considerations used for that decision.

As a reminder, an ellipse has five degrees of freedom. One parameterization is the length of the major and minor semi-axes a and b, the center point (x_0, y_0) and a rotation angle γ around the center point. The equation for an ellipse with these

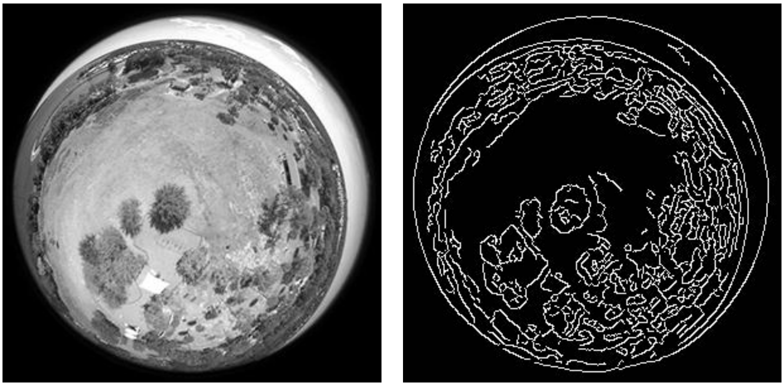

Figure 5.3: Fisheye image and corresponding edge map.

parameters is

$$(\frac{x-x_0}{a}\cos\gamma+\frac{y-y_0}{b}\sin\gamma)^2+(\frac{y-y_0}{b}\cos\gamma-\frac{x-x_0}{a}\sin\gamma)^2=1 \qquad (5.2)$$

The camera attitude can be expressed either as the pitch and roll angles, θ and ϕ, or a tilt angle α around a rotation axis n_{rot}. For a fisheye camera with its optical axis aligned with the gravity vector (vertical), the projection of the horizon onto the unit sphere will be a horizontal circle. The radius r of the circle is determined by the camera altitude h. The radius of the horizon circle on the unit sphere will remain constant irrespective of the camera attitude. As seen in the above Hough circle transform example, being able to simply reduce the voting space one dimension from one input parameter, in this case the altitude h, is very advantageous.

For the vertical camera, the image of the horizon is a circle, i.e. the length of the two semi-axes a and b are the same. When tilting the camera, the major axis of the ellipse will be parallel to the rotation axis n_{rot}. The length a of the major axis will increase with the tilt angle α whereas the length b of the minor axis will decrease with the tilt angle, i.e. the length of the two principal axes will vary both with the altitude h and the tilt angle α. These constraints do not cater for an easy reduction of the voting space.

Further, we define the edge direction in a point p on the image plane as $(-\nabla_y, \nabla_x)$, i.e. normal to the gradient direction. When projecting a point and its edge direction onto the unit sphere, the projected edge direction will constitute a tangent vector for the plane of the horizon circle in that point, see figure 5.4.

But if we have the projected point P on the unit sphere, a tangent vector $\mathbf{t}$ for the horizon circle, and we also know the radius $r(h)$ for that circle, the normal vector $\mathbf{n}$ for the horizon circle plane can be readily computed. The sought pitch and roll angles are then easily deduced from the normal vector direction. The explicit equations for these computations are given in papers B and C.

On the image plane, the edge direction sets a constraint on the ellipse parameters which is obtained by differentiating (5.2). The equations and interpretation

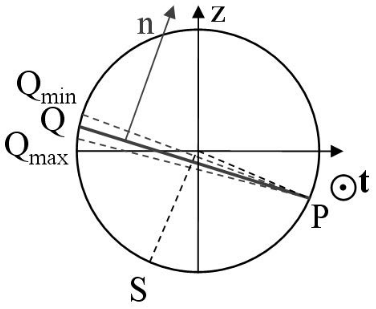

Figure 5.4: Estimate of horizon normal **n** from edge points. Tangent vector **t** is directed out of paper.

are far from being as trivial as for the tangent vector on the unit sphere. A second constraint is that the edge point shall lie on the ellipse and satisfy (5.2). Combined with the previous three constraints related to the length and direction of the ellipse axes, there is a total of five constraints which is sufficient to compute an estimate of the attitude angles. However, the mathematics for computing the attitude angles from the constraints given by the ellipse shape on the image plane is truly intricate and not a tractable solution to the attitude estimation problem. This is the main reason for the decision to perform the Hough voting on the unit sphere and not on the image plane.

5.3 Hough voting - considerations for real images

For real world images, the horizon does not generate a perfect ellipse on the image plane and the circle shape on the unit sphere is also an approximation. It is mainly the true image scene content that causes the shape imperfection, but also image noise and camera model errors contribute to this deviation from the ideal shape. What considerations need to be taken to accomodate for these imperfections in the Hough voting?

In papers B and C where the voting was carried out over a pitch and roll angle accumulator array, the shape and gradient direction imperfections imply that the computed attitudes for the different edge pixels will be spread out in a neighborhood around the true attitude value. The aggregated votes may create local maxima in the accumulator array and just picking the cell with the maximum score may lead to an erroneous estimate. Therefore, it is common to filter the accumulator array with a smoothing kernel prior to extracting the cell with the maximum value.

5.4 Extraction of horizon edge pixels

In paper B, the horizon was detected in fisheye images using a Canny edge detector and a probabilistic Hough voting scheme as described in section 5.2. The horizon detection and the attitude estimate were based on the assumption that the earth was a smooth sphere at sea level with no topography. Figure 5.5 a)-b) shows an example image and the corresponding edge map used in the voting. The estimated horizon is marked white in the edge map.

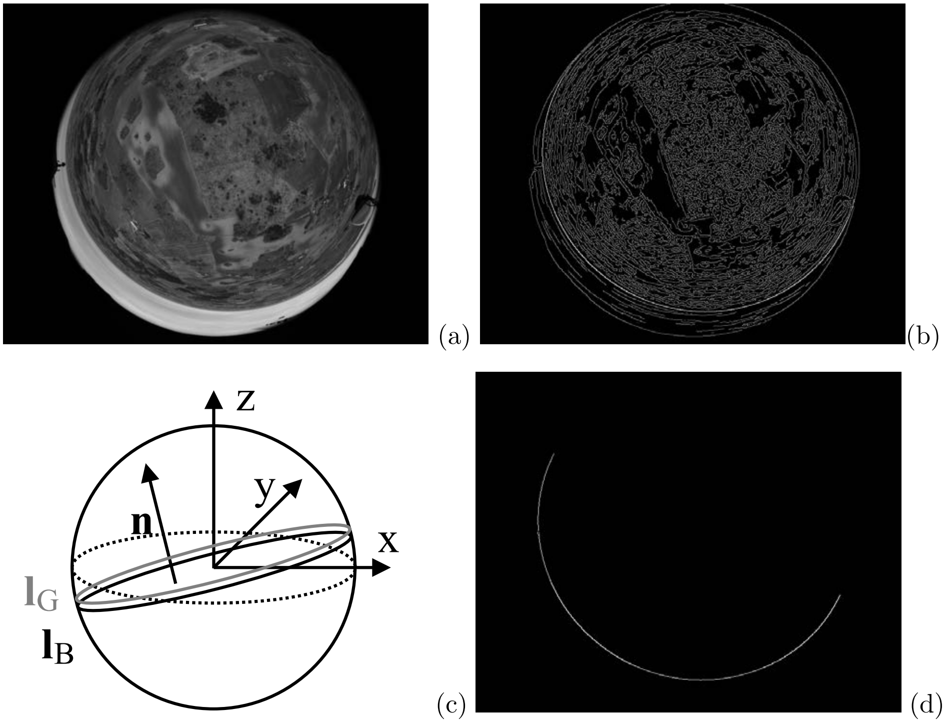

Figure 5.5: (a) Original image. (b) Canny edge map with estimated horizon from Hough voting in white. (c) Extract edge pixels in band between lines l_B and l_G. (d) Extracted horizon edge pixels.

In paper C, the attitude estimate was refined by registration of the detected horizon with the geometric horizon from a DEM. To perform the refinement, we need to extract only the edge pixels originating from the horizon in the image. This is done geometrically by reasoning as follows. For a perfectly calibrated camera and exact knowledge of the camera altitude, the ellipse on the image plane corresponding to the estimated horizon from the Hough voting will always be slightly smaller than the true perceived horizon on the image due to the topography on top of the ideal spherical earth. Thus, most of the true horizon edge pixels will be on or outside the estimated horizon. Due to quantization effects in the edge detector (only integer pixels), some true horizon edge pixels may be 1/2 pixel inside the estimated horizon. If the shift of the horizon due to topography is less than one pixel, it is sufficient to project the estimated horizon on the image plane and extract all edge pixels that are within a 3x3 matrix from the horizon pixels on the image plane.

For high resolution images, and when the ground elevation in the scene is large, the shift of the horizon due to topography may be larger than one pixel. For the flight in paper C, we could use DEM data (highest altitude in the area) to compute an upper angular limit for the shift of the horizon to 0.4° and denote it β_{lim}. This means that all true horizon pixels on the image will be projected onto the unit sphere in a thin band above the estimated horizon as given by the probabilistic Hough voting, figure 5.5 c). The black line $\mathbf{l}_{\text{B}}$ in the figure is the estimated horizon. The gray line $\mathbf{l}_{\text{G}}$ is generated by points that make an angle β_{lim} with the horizon points. Explicit equations for the lines are given in paper C.

We project the band between the lines $\mathbf{l}_{\text{B}}$ and $\mathbf{l}_{\text{G}}$ onto the image plane to create a mask for potential horizon edge pixels. Since the edge detector gives integer pixel values, we include a 3x3 neighborhood around each projection point on the image plane in the mask. From the Canny edge image, we only extract the edge pixels within the mask for the subsequent attitude refining process. The extracted horizon edge pixels, figure 5.5 d), are used in the subsequent registration with the geometric horizon for attitude estimate refinement.

5.5 Hough voting - leaf detection

During the course of the studies, another application involving ellipse detection was encountered. The GARNICS project [10] aims at 3D sensing of plant growth. One specific task is to automatically monitor and measure the growth of leaves on tobacco plants. The leaves on these plants can be well approximated with an ellipse extending from the stalk, see figure 5.6.

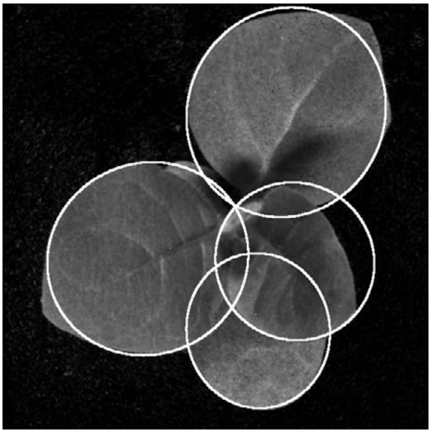

Figure 5.6: Tobacco plant with estimated leaf size from Hough voting scheme.

A slightly modified version of the Hough voting scheme used for horizon detection, described in section 5.2, was employed to detect and estimate the size of the tobacco leaves.

There are two major differences detecting the leaves compared to the horizon detection. If no prior information is given, the number of leaves in an image and the size of the leaves are both unknown. The unknown size corresponds to an unknown radius when projecting the leaf edge pixels onto the unit sphere. The position of a leaf in an image, relative to the stalk, corresponds to the pitch and roll angles on the unit sphere and in parameter space. In total, three parameters are required to describe a leaf on the unit sphere and they are all unknown, i.e. the accumulator array needs to be three dimensional. The unknown number of leaves in an image also implies that we need to search for and accept local minima in the accumulator array as true objects. This is in contrast to the horizon detection where the horizon generated a global maximum in the accumulator array.

The result obtained estimating the size of tobacco leaves in one example image using the modified Hough voting scheme is shown in figure 5.6. The result also illustrates that the method developed for horizon detection could be used for other applications involving ellipse detection.

Chapter 6

Vision-based global pose estimation

The output from the vision-based methods developed in this thesis is a global vehicle pose estimate. *Global* in this context means that the estimated pose is given in a world coordinate frame, e.g. the WGS84 or SWEREF99 coordinate systems. To achieve the aim with global pose estimation, we must somehow find a relationship between the image content and geographic information given in a world coordinate frame. This process is called *image registration*, i.e. transforming different sets of data into one common coordinate system.

6.1 Image registration

The most appropriate image registration method for a given set of aerial images to infer the global pose will of course not only depend on the camera type and the image scene content but also on the geographic information available. The latter could be digital elevation models or textured 3D models, as described in section 2.2, but it could also be 3D point clouds from Lidar measurements or a large database with images captured at known positions and orientations.

The sections below will introduce the concept of some conceivable image registration techniques for aerial images and describe the main principle for the global pose estimation methods proposed in papers A, B and C.

6.2 3D-2D registration

In paper A, the problem formulation assumed that we had a textured 3D model at hand, generated from aerial images with Saab's Rapid 3D Mapping$^{\text{TM}}$ process [32]. The 3D model was generated from images captured in the Linköping area in the summer 2008. The query images, for which the camera pose was to be estimated, were captured in March 2010 when there was still some snow on the ground. This is a common situation, the geographic information (3D model) and

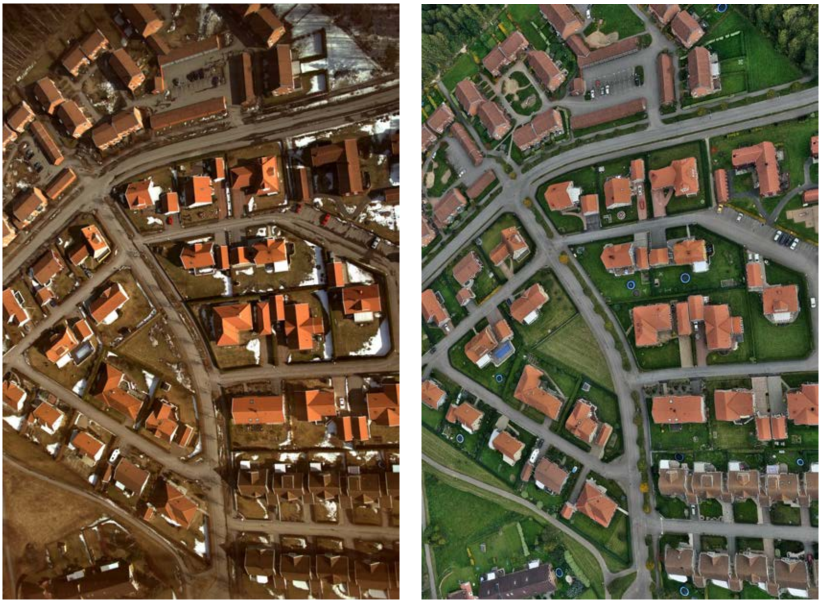

Figure 6.1: True and rendered aerial image from the same camera pose.

the query images are from different instances in time and the registration method must be able to cope with season variations affecting the data.

A commonly used technique for image registration when a 3D model of the area is available is to assume a 6DoF pose for the query image and given this assumption render an image from a virtual camera in the same pose in the 3D model. The two images are then to be registered which is called 3D-2D registration. In our case for paper A, how would a rendered image from the 3D model and a true image from the flight compare? This is illustrated in figure 6.1. Note the presence of snow and the brownish ground in the query image and the difference in foliage on the trees along the north-south bound street in the center of the images.

If these two images are registered, e.g. by establishing image correspondences, the relative rotation and translation between the query image and the rendered image can be estimated using the epipolar constraint. Based on this estimate, a refined camera pose for the query image can be computed, and a new image is rendered for renewed registration. This process is repeated until the pose update is small and below a predefined threshold. The process requires a good initialization for the camera pose to converge to the true global pose.

What methods are available to find image correspondences between a true image and a rendered image from different seasons? Intensity based and feature based methods may be used but are likely to generate a large fraction of image correspondence mismatches. Another method is proposed in [22] where mutual information was successfully used to register aerial images from an urban scene with rendered images from a 3D Lidar model.

6.3 3D-3D registration

Paper A - registration of height information

To obtain a method that is robust even in the presence of considerably more snow than in figure 6.1, the pose estimation method proposed in paper A is based on registration of height information computed from images and extracted from a DEM, i.e. a 3D-3D registration.

A dense 3D reconstruction was generated using motion stereo with the concept described in section 4.7. A rough global pose was assumed for camera 1. The relative rotation between the camera poses was taken from IMU data and the relative translation direction was estimated from FAST image correspondences and the epipolar constraint. The dense 3D reconstruction was then orthorectified based on the global pose assumption for camera 1 to generate a local height patch, i.e. the height h_p for the local area jointly seen in the two images. A height map h_m is extracted from the 3D-model from, what is assumed to be, the corresponding area. Example images for the local height patch and the corresponding height map are shown in figure 6.2. The intensity in the images represent height above the ground.

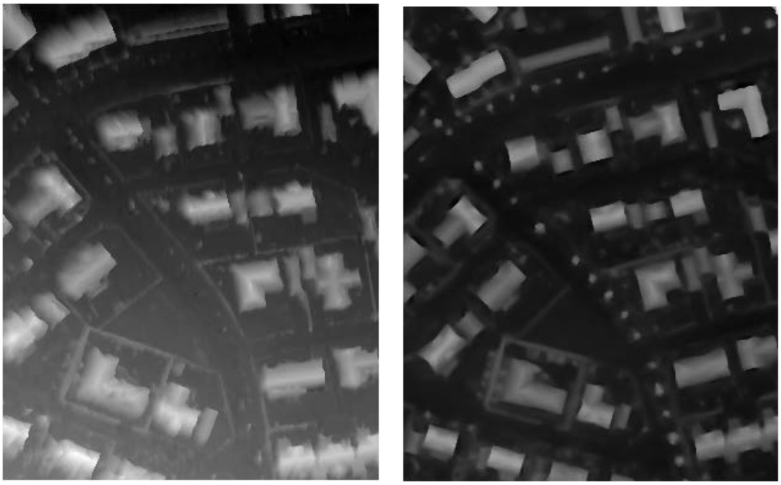

Figure 6.2: Local height patch (left) and 3D map (right) from the same urban area.

Due to the errors in the global pose assumption for camera 1, there will be a rotation $\mathbf{R}$ and a translation $\mathbf{t}$ between the height patch and the height map. Further, the absolute scale S for the height patch is not accurately known. In total, the mapping between a 3D point in the height patch $\mathbf{x}_\mathrm{p}$ and the corresponding point in the height map $\mathbf{x}_\mathrm{m}$ is given by

$$\begin{pmatrix} x_\mathrm{p} \\ y_\mathrm{p} \\ z_\mathrm{p} \\ 1 \end{pmatrix} = \begin{pmatrix} \mathrm{S}\mathbf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_\mathrm{m} \\ y_\mathrm{m} \\ z_\mathrm{m} \\ 1 \end{pmatrix} \tag{6.1}$$

A detailed comparison of the height information in figure 6.2 shows that the height contour of the houses is more diffuse in the local height patch than in the height map. The effect is enhanced at remote house boundaries and is caused by stereo shadowing effects for the height patch which is based on two images only. This type of difference is not accounted for in (6.1) and is one error source for non-perfect registration.

Registration of two surfaces in 3D space could be solved e.g. using ICP (Iterative Closest Point) [4], but compared to available methods for ICP in the literature, the scale is unknown in our case. The proposed solution in paper A is to split the 3D registration into solving three 2D subproblems; first registration in the ground plane (XY), then in the two vertical planes (XZ) and (YZ). The justification for this approach is that the largest errors in our specific problem formulation, the XY position, the heading angle and the scale, are all considered in the ground plane, i.e. in the first 2D subproblem. The errors in the pitch and roll angles are expected to be considerably smaller, either using an estimate obtained from the methods in papers B or C or from an onboard IMU.

For alignment in the different planes we use an approach similar to the tracking part in [30]. We minimize the error function ϵ_{IJ}, the squared difference in height between the local height patch $h_\mathrm{p}(\mathbf{x})$ and the map $h_\mathrm{m}(\mathbf{x})$,

$$\epsilon_{IJ} = \int_W \left[h_\mathrm{p}(\mathbf{A}\mathbf{x} + \mathbf{d}) - h_\mathrm{m}(\mathbf{x})\right]^2 \omega(\mathbf{x})\, d\mathbf{x} \ , \tag{6.2}$$

where $\mathbf{A}$ denotes a linear transformation, $\mathbf{x}$ is a point in the image, $\mathbf{d} = [d_\mathrm{x} \ \ d_\mathrm{y}]^T$ is a displacement and ω is a window function. The indices IJ denote the three different planes. The integration domain W is the whole patch area.

The error function in (6.2) is minimized by linearizing the integrand and differentiating with respect to all unknowns. In the XY-plane, the matrix $\mathbf{A}$ was a general linear transform whereas in the two vertical planes it was constrained to be a rotation matrix. The explicit error functions to be minimized in the different planes are given in paper A.

From the alignment in the three different planes, an estimate of the rotation, translation and scale in (6.1) is obtained and from this information a refined estimate of the global pose can be computed.

Papers B and C - registration of 3D points on the unit sphere

The global pose estimation method in paper A requires a good initialization of the absolute pose for camera 1 to converge to a pose close to the true value. If the pitch and roll angles are known with high accuracy, the 3D registration problem in paper A would collapse into a 2D registration problem in the ground plane.

This reasoning was one of the drivers prior to developing the methods for attitude estimation in papers B and C.

For an aircraft flying at a few hundred meters altitude, the horizon can most often be seen and used as an attitude guide. Of course, haze, fog, rain and snow and other weather phenomena can place restrictions, but this is a known limitation for visual sensors. The use of a fisheye camera with more than 180° FOV mounted underneath the aircraft was considered a tractable sensor to infer the complete 6DoF pose. The horizon in the image could be used for attitude estimation, and the part of the image viewing the ground could be used for global pose estimation, e.g. using the method in paper A.

The method proposed in paper B, described in section 5.2, used edge detection and probabilistic Hough voting to estimate the aircraft attitude assuming a smooth spherical earth. The 3D-3D registration is here implicit in the process. We exploit the fact that for a smooth earth with no topography the horizon will generate a circular disc with radius $r(h)$ on the unit sphere. This is the connection between image content and the world reference frame that enables a global attitude estimate to be provided. The attitude errors using this method is from a few tenths of a degree up to one degree in an area with modest ground elevations, i.e. up to a few hundred meters. This is definitely acceptable as an initialization for the method in paper A, but not accurate enough for the complete 6DoF pose estimation to be considered a mere 2D registration problem.

In paper C, all edge pixels within a thin band around the ideal horizon were extracted and projected onto the unit sphere. This point set is denoted $\mathbf{P}_{\mathrm{s}}$. If this point set is rotated with the transpose of the estimated camera rotation $\mathbf{R}_{\mathrm{c,est}}$, the new rotated point set on the sphere $\mathbf{P}_{\mathrm{r}}$ will ideally be the perceived horizon points rotated to a camera system aligned with the world coordinate system. Given an assumption of the aircraft global position and heading, a geometric horizon can be computed from a DEM and be projected onto the unit sphere. This point set from the geometric horizon is denoted $\mathbf{P}_{\mathrm{geom}}$. We then have two point sets on the unit sphere, one originating from the image and one from the DEM. An example is shown in figure 6.3.

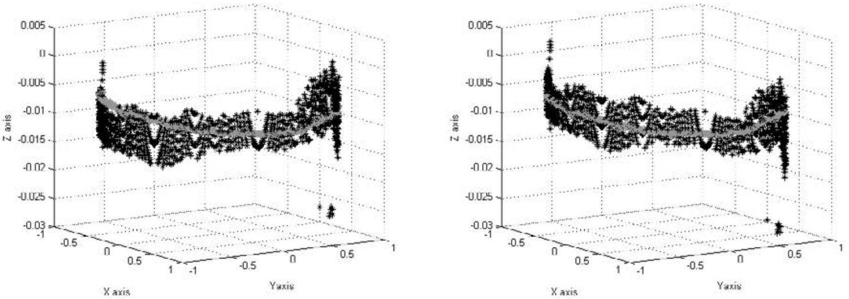

Figure 6.3: Horizon from image edge pixels (black) and from DEM (gray). a) without refinement, b) after refinement. Note the different scale for the Z-axis.

To refine the attitude estimate $\mathbf{R}_{\mathrm{c,est}}$ obtained when assuming an ideal horizon, we search for a rotation such that the distance between the corresponding points in the two point sets is minimized. If we choose to compute the point set $\mathbf{P}_{\mathrm{geom}}$ for the same angles α_i on the XY-plane as for the point set $\mathbf{P}_{\mathrm{r}}$, i.e.

$$\alpha_i = \tan^{-1}(\mathrm{P}_{\mathrm{r_y},i}/\mathrm{P}_{\mathrm{r_x},i}) \tag{6.3}$$

the small arcs between the point sets will have their main component along the z-axis of the unit sphere. Using a small arc angle approximation, we define the error function to be the squared distance between the z-coordinate for the point sets, i.e.

$$\epsilon_z = \sum_i (\mathrm{P}_{\mathrm{r_z},i} - \mathrm{P}_{\mathrm{geom_z},i})^2 \tag{6.4}$$

where the summation is made over all extracted horizon edge points.

For attitude refinement, we search for the camera orientation $\mathbf{R}_{\mathrm{c}}$ that minimizes the error function in (6.4), i.e.

$$\mathbf{R}_{\mathrm{c,est}} = \arg\min_{\mathbf{R}_{\mathrm{c}}} \epsilon_z(\mathbf{R}_{\mathrm{c}}). \tag{6.5}$$

We minimize the nonlinear error function using the Levenberg-Marquardt algorithm and initialize with the camera orientation matrix obtained from the Hough voting.

Another conceivable solution to minimize the distance between the two point sets would be the orthogonal Procrustes method [29].

Chapter 7

Evaluation

The aim with the methods presented in this thesis was to estimate the global pose of the airborne vehicle carrying the camera. The most relevant error measure for the results hence is a comparison with the ground truth global pose if available. One big advantage being a student and at the same time being employed by the aircraft manufacturer SAAB is the availability of true aerial images with accurate ground truth data. Having a ground truth global pose for aerial imagery is relatively rare in academia which has become evident when comparing the performance of the developed methods with other methods presented in the literature. It is very common that subjective measures are used when evaluating how well images are registered with a 3D model. Expressions like ”a good visual match with the model was obtained” are frequently encountered and most often no absolute errors are presented.

7.1 Ground truth generation

In paper A, two large sets of aerial images from the Linköping area were used to evaluate the accuracy of the proposed method. Both sets of images had been processed by SAAB to construct a 3D-model using their Rapid 3D Mapping process. The ground truth pose used for the individual images is the output from the bundle adjustment in this process. Evaluation of these 3D-models has shown that the accuracy of the global position for landmarks on the ground is better than 1 dm. To achieve this accuracy when flying at 600 m altitude, the attitude accuracy for the euler angles (yaw, pitch and roll) must be better than 0.01° and the camera position accuracy a few centimeters.

For paper C, a dedicated flight with a fisheye camera was performed to capture the images required in order to evaluate the performance of the global vehicle attitude estimates. Onboard the aircraft, there was a GPS and a highly accurate IMU. A filtered GPS/IMU solution from the flight was computed by the navigation department at SAAB. The position accuracy is claimed to be a few centimeters and the attitude errors about 0.01° for pitch and roll and 0.04° for the yaw angle.

7.2 Evaluation measures

In paper A, the global 6DoF pose for a pair of images was estimated. The method is based on registration of a local height patch, computed with motion stereo, with a 3D model. The distance travelled for the camera between the two images is also unknown and estimated as a scale ratio between the local height patch and the 3D model. For the 6DoF pose, the deviation from the true global position and the deviation in the vehicle euler angles (yaw, pitch and roll) was used as the error measure for the method. For the scale, i.e. the stereo baselength between the cameras, the relative error in percent is used as the evaluation measure.

In paper A, the registration rate as a function of the initialization errors was reported to compare the performance of our method with similar methods in the literature. The judgement whether a registration is considered correct or not is to some degree a subjective measure. Just by plotting the absolute 6DoF pose errors, it was very evident which pose estimates that had been trapped in a local minimum in the optimization process and generated large errors. For a few cases, where the absolute orientation errors were in a grey zone, a visual inspection was made to judge if the height patch and the height map were well aligned. Figure 7.1 shows an example where the local height patch, after transformation with the estimated pose, is well registered with the height map.

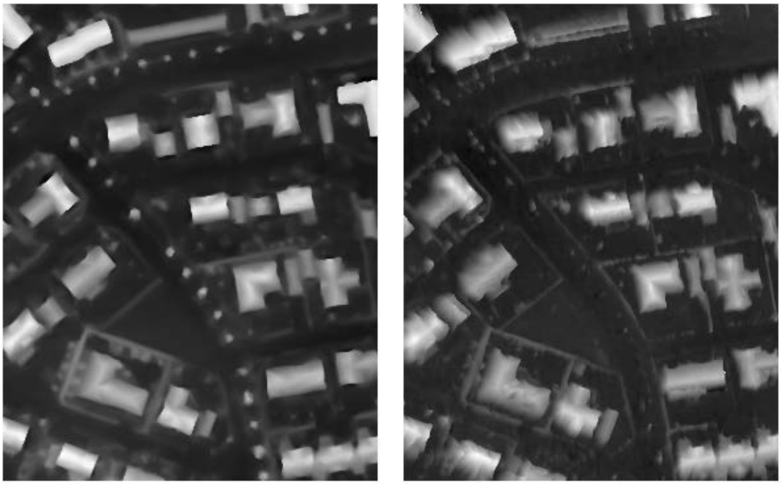

Figure 7.1: Height map from model (left). Local height patch after transformation with estimated pose (right).

For paper B, aerial images from the internet were used to evaluate the developed horizon estimation method. For these images there was no ground truth pose and no camera calibration available. To evaluate the performance of the method, a comparison was made between the manually estimated horizon and the horizon computed by the method. Two images from paper B are shown in figure 7.2 where the estimated horizon is overlaid on the image.

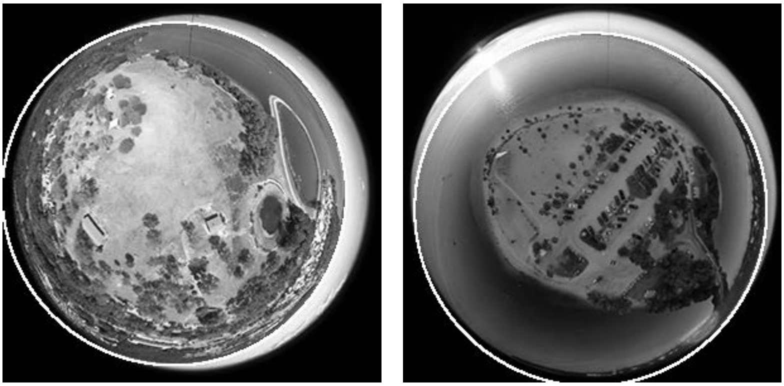

Figure 7.2: Images from paper B with the estimated horizon overlaid. Images are from markmarano.com

In paper C, the absolute global pose for the aircraft was available from navigation sensors. Our method estimates the absolute pitch and roll angles via horizon detection and the evaluation measure is the deviation in the pitch and roll angles from the ground truth value. Two images from paper C are shown in figure 7.3 where the estimated horizon is overlaid on the image.

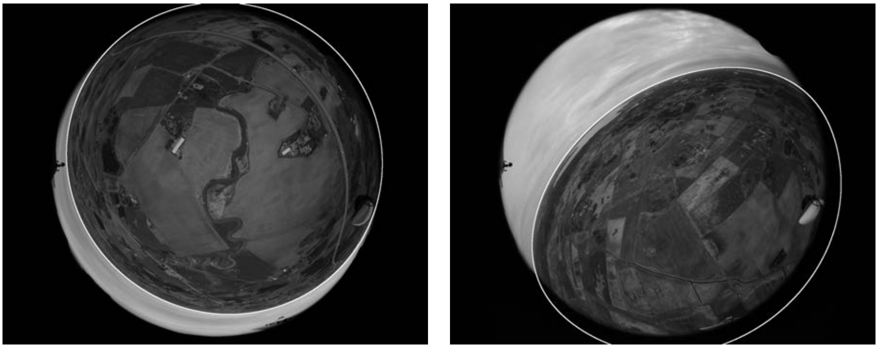

Figure 7.3: Images from paper C with the estimated horizon overlaid.

Chapter 8

Concluding Remarks

The aim of this thesis has been the development of vision-based methods for global pose estimation of airborne vehicles for ego-positioning and to aid their navigation. It is obvious that no single method can be used for pose estimation in all different phases in a flight and in all conceivable scenarios. The image content will be very different on the runway, during ascent, during flight at low or high altitude, above urban or rural areas, flights performed in different seasons, and so on. In total, a multitude of methods will be required to handle all these situations. With this total scope in mind, the methods and the results presented in this thesis has contributed with a few more pieces to solve the overall large jigsaw puzzle, automated vision-based methods for ego-positioning and navigation of an airborne vehicle.

8.1 Conclusions of results

Paper A presented a method for global pose estimation in an urban area assuming that the vehicle is flying at a sufficient height above the houses and treetops to enable the reconstruction of a local height map of the area. The local height map is registered with a georeferenced 3D model of the area to infer the global pose. The reason for matching altitude information is to make the method season invariant which is essential in the Swedish climate where images of the same area may look very different in various seasons. The method in paper A is more robust than any previously reported method on similar datasets in the sense that it provides the same, high level of correct registration rate for larger initialization errors than other methods. A fair comparison with other methods regarding the accuracy of the global pose estimate cannot be made since most other methods do not have a ground truth pose available for their images. A limitation with the method is that it requires altitude structure in the scene and that it also needs a fairly good initialization not to be trapped in a local minimum in the optimization process.

The method for attitude estimation (pitch and roll angle) in paper B is one of only a few methods in the literature that use an omnidirectional (fisheye) camera for horizon detection. It is the first paper on fisheye images employing a probabilistic Hough voting scheme for horizon detection which enables the probability density functions for the altitude and the pitch and roll angles to be used to make the attitude estimate more robust and suppress other edge pixels in the image. Since the method in paper B was evaluated on fisheye images from the internet with no ground truth pose and camera calibration available, no quantitative statements concerning the attitude accuracy could be made. However, the method proved promising enough to continue with the concept and perform our own flight trials with a fisheye camera to further evaluate the method.

The method in paper C builds on the results and the method in paper B. The additions made are unique and it is the very first time that an attitude estimate from horizon detection in omnidirectional images is refined through registration with the geometrically expected horizon from a digital elevation model. In a flight with more than 1600 images, the mean and standard deviation for the pitch and roll angle errors are smaller than 0.02° and 0.05° respectively. These errors are roughly one order of magnitude smaller than the errors reported for previous methods on aerial images. To achieve the highly accurate attitude estimates, the ray refraction in the earth atmosphere was taken into account. This component is essential for highly accurate attitude estimates but has not been included in any previously reported attitude estimation method based on horizon detection.

8.2 Future work

As stated above, many more methods for vision-based pose estimation are required before vision alone can be robustly used throughout a flight for ego-positioning and autonomous navigation of an airborne vehicle. The use of omnidirectional images for this purpose is surprisingly rare in the work presented in the literature. My belief is that there is a huge potential for this type of images in the applications mentioned above, thanks to the very large field of view. For instance, one part of the fisheye image viewing the horizon could be used for attitude estimation using the method in paper C. Another part of the fisheye image, viewing the ground, could be used as in paper A to estimate the full global 6DoF pose.

During the starting phase of the flight, a belly-mounted fisheye camera will still be able to view some of the surroundings in the periphery of the image. This information could conceivably be used in a visual odometer to maintain a good pose estimate until the vehicle has ascended where other vision-based methods can take over if provided with a good initialization.

Bibliography

[1] D. H. Ballard. Generalizing the Hough Transform to Detect Arbitrary Shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[2] G. Bao, Z. Zhou, S. Xiong, X. Lin, and X. Ye. Towards Micro Air Vehicle Flight Autonomy Research on The Method of Horizon Extraction. *IMTC*, 2003.

[3] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *ECCV*, pages 404–417, 2006.

[4] P. J. Besl and N. D. McKay. A method for registration of 3D shapes. *Pattern Analysis and Machine Intelligence*, 1992.

[5] J. Canny. A computational approach to edge detection. *PAMI*, 8:679–698, 1986.

[6] C. Demonceaux, P. Vasseur, and C. Pégard. Omnidirectional vision on UAV for attitude computation. *International Conference on Intelligent Robots and Systems*, 2006.

[7] S. J. Dumble and P. W. Gibbens. Horizon Profile Detection for Attitude Estimation. *Journal of Intelligent Robotic Systems*, 68:339–357, 2012.

[8] D. Dusha, L. Mejias, and R. Walker. Fixed-Wing Attitude Estimation Using Temporal Tracking of the Horizon and Optical Flow. *Journal of Field Robotics*, 28(3):355–372, 2011.

[9] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *In Communications of the ACM*, 24(6):381–395, 1981.

[10] http://www.garnics.eu.

[11] G. Granlund and H. Knutsson. Signal Processing for Computer Vision. *Springer, ISBN 0-7929-9530-1*, 1995.

[12] B. Grelsson and M. Felsberg. Probabilistic Hough Voting for Attitude Estimation from Aerial Fisheye Images. *18th Scandinavian Conference in Image Analysis, SCIA*, pages 478–488, 2013.

[13] B. Grelsson, M. Felsberg, and F. Isaksson. Efficient 7D Aerial Pose Estimation. *IEEE Workshop on Robot Vision*, 2013.

[14] B. Grelsson, M. Felsberg, and F. Isaksson. Global Pose Estimation of Aerial Images. *SSBA*, 2013.

[15] B. Grelsson, M. Felsberg, and F. Isaksson. Highly accurate attitude estimation via horizon detection. *Submitted to Journal of Field Robotics*, 2014.

[16] M.S. Gyer. Methods for Computing Photogrammetric Refraction Corrections for Vertical and Oblique Photographs. *Photogrammetric Engineering and Remote Sensing*, 62(3):301–310, 1996.

[17] R.I. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. *Cambridge University Press, 2nd edition*, 2004.

[18] P. Hough. Method and means for recognizing complex patterns. *U.S. Patent 3069654*, 1962.

[19] K. Levenberg. A method for the solution of certain problems in least squares. *Quarterly of applied mathematics*, 2:164–168, 1944.

[20] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 20(2):91–110, 2003.

[21] B.D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. *In Proc. of Imaging Understanding Workshop*, 1981.

[22] A. Mastin, J. Kepner, and J. Fisher. Automatic Registration of LIDAR and Optical Images of Urban Scenes. *CVPR*, 2009.

[23] T. G. McGee, R. Sengupta, and K. Hedrick. Obstacle Detection for Small Autonomous Aircraft Using Sky Segmentation. *ICRA*, 2005.

[24] C. Mei and P. Rives. Single View Point Omnidirectional Camera Calibration from Planar Grids. *IEEE International Conference on Robotics and Automation*, 2007.

[25] http://www2.jpl.nasa.gov/srtm/.

[26] D. Nister. An efficient solution to the five-point relative pose problem. *Proc. Computer Vision and Pattern Recognition*, pages 195–202, 2003.

[27] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. *ECCV*, 2006.

[28] D. Scaramuzza, A. Martinelli, and R. Siegwart. A toolbox for Easily Calibrating Omnidirectional Cameras. *International Conference on Intelligent Robots and Systems*, 2006.

[29] P. H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[30] J. Shi and C. Tomasi. Good features to track. *CVPR*, 1994.

[31] S. Thurrowgood, D. Soccol, R. J. D. Moore, D. Bland, and M. V. Srinivasan. A Vision Based System for Attitude Estimation of UAVs. *International Conference on Intelligent Robots and Systems*, 2009.

[32] `http://www.saabgroup.com/vricon`.

[33] N. J. Wade and S. Finger. The eye as an optical instrument: from camera obscura to Helmholtz's perspective. *Perception*, 30(10):1157–1177, 2001.

[34] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

Part II

Publications

Publications

The articles associated with this thesis have been removed for copyright reasons. For more details about these see:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-108213