Saeed Yahyanejad, M.Sc.

Orthorectified Mosaicking of Images from Small-scale Unmanned Aerial Vehicles

DISSERTATION

zur Erlangung des akademischen Grades
Doktor der Technischen Wissenschaften

Alpen-Adria Universität Klagenfurt
Fakultät für Technische Wissenschaften

ALPEN-ADRIA
UNIVERSITÄT
KLAGENFURT | WIEN GRAZ

Institut für Vernetzte und Eingebettete Systeme

1. Begutachter: Univ.–Prof. Dr. techn. Bernhard Rinner
Institut für Vernetzte und Eingebettete Systeme,
Alpen-Adria Universität Klagenfurt

2. Begutachter: Univ.–Prof. Gian Luca Foresti
Department of Mathematics and Computer Science,
University of Udine

Klagenfurt, March 2013

Declaration of Honor

I hereby confirm on my honor that I personally prepared the present academic work and carried out myself the activities directly involved with it. I also confirm that I have used no resources other than those declared. All formulations and concepts adopted literally or in their essential content from printed, unprinted or Internet sources have been cited according to the rules for academic work and identified by means of footnotes or other precise indications of source.

The support provided during the work, including significant assistance from my supervisor has been indicated in full.

The academic work has not been submitted to any other examination authority. The work is submitted in printed and electronic form. I confirm that the content of the digital version is completely identical to that of the printed version.

I am aware that a false declaration will have legal consequences.

(Signature) (Place, Date)

Abstract

Unmanned aerial vehicles (UAVs) have been recently deployed in various civilian applications such as environmental monitoring, aerial imagery, and surveillance. Small-scale UAVs are of special interest for first responders since they can rather easily provide bird's eye view images of disaster areas. For such UAVs the number of images and the positions where to capture them are predefined due to limitations in flight time, communication bandwidth and local processing. The main goal of this thesis is to develop methods for mosaicking of individual aerial images taken from homogeneous or heterogeneous sensors on small-scale UAVs. The mosaicking of images taken in such scenarios are challenging as compared to panoramic construction or other mosaicking methods such as satellite image mosaicking. When flying with UAVs at a relatively low altitude (below 100m), non-planar objects on the ground make the feature matching and image registration more difficult. In addition, other artifacts such as dynamic scene, lens distortion, and heterogeneous sensors makes the mosaicking procedure more difficult.

In this thesis we focus on producing orthorectified and incremental mosaics from low-altitude aerial images. The orthorectification is important in order to preserve the relative distances in the mosaic. On the other hand, the incremental mosaicking means to update the real-time mosaic while individual images are being added. We present two methods to construct such mosaics. The first method combines the metadata of the images such as GPS positions and the UAV orientations with the image processing techniques to construct the mosaic. The second method does not exploit any metadata and only uses the images. By this method we find and mitigate the sources of errors, in the process of incremental mosaicking, to achieve an orthorectified mosaic. Unlike some other mosaicking approaches we avoid any global optimization because of the high computational complexity. Furthermore, the global optimization methods require all images at once while in our incremental mosaicking we do not reposition any of the the previously mosaicked images.

Eventually we demonstrate some novel methods for multispectral aerial imagery with thermal and visual (also referred to as RGB) cameras. We show how to register the images of different spectrums and how to improve the quality of this interspectral registration. The contribution of this part includes (i) the introduction of a feature descriptor for robustly identifying correspondences in images of different spectrums, (ii) the registration of image mosaics, and (iii) the registration based on depth maps.

Acknowledgments

First of all I would like to express my immense appreciation to Professor Bernhard Rinner, my supervisor who played a large role in shaping my thesis. Needless to say, without his guidance, encouragement, and tremendous patience I would not be able to accomplish this project on my own.

I would also like to acknowledge the financial support of the project *Collaborative Microdrones (cDrones)* in the research cluster Lakeside Labs which was partly funded by the European Regional Development Fund, the Carinthian Economic Promotion Fund (KWF), and the state of Austria under grant KWF-20214/17095/24772.

I am also grateful to all my colleagues and friends at the University of Klagenfurt and especially the Institute of Networked and Embedded Systems (NES) and the cDrones team for their continuous and endless support. You have never hesitated to help me both scientifically and technically which has been a truly rewarding experience.

Finally, yet importantly, I take the opportunity to thank my beloved parents and sister for their constant love, concern and encouragement throughout my life and education. In addition, I am very privileged to get to know Viktoria and Patrick during this period, who had a profound effect on me and my family.

Contents

List of Figures

List of Tables

CHAPTER 1 Introduction

In recent years, images have become extremely important in our daily life. No matter whether presenting in a conference, broadcasting news or writing on web-pages, it bores people when there is no image to visualize the concepts. Although vision is the most important sense of human in terms of range and speed, the brain is not able to process all the visual information and details in a glance. Human eyes see a continuous chain of events, while a captured image freezes a moment in time. Most of us have heard the adage "a picture is worth a thousand words". It means that images can contain and convey complex concepts and information. They are used to explain different phenomena or events, since we can study and process them later in time. Thus, images are being used ubiquitously in different fields such as mobile phones, medical (e.g., X-ray and microscopic images), remote sensing (e.g., aircraft and satellite images), astronomy (e.g., telescope images), underwater photography, art, and advertisement.

In this thesis we focus on a specific field of imagery, called *aerial imagery*, with many applications such as management of natural disaster, monitoring the environment, and surveillance. The first aerial images were taken from balloons, however, with advances of technology the aerial imagery techniques have exploited advanced aircrafts. In our research, we use small-scale unmanned aerial vehicles (UAVs) to take aerial images from low altitudes (below 100m). In this way, we provide a recent information about a target area, with a relatively low cost. Capturing images from low altitude provides more detailed information of a target area. Flying in low altitudes, on the other hand, needs more individual images to cover a large area. In case of multiple images, it is easier to extract the required data if we first combine the information from different images. One way of combining these information is to align or mosaick them together. Image mosaicking have been in practice since long ago, before the invention of digital images [74]. Initially, it started by manually aligning the images like pieces of a puzzle. The need for mosaicking (e.g., constructing topographic maps) expanded over time with the advent of digital images and satellite imagery. Eventually, more sophisticated and reliable mosaicking methods came along to automatically produce image mosaics.

In our work we survey the process of aerial image mosaicking and present different methods to construct an accurate mosaic. This mosaic is comprised of individual images taken from low-altitude UAVs with different sensors. It is intended to provide the user the desired information regarding a target area. Google Maps [1] and Bing Maps [2] are samples that provide similar mosaics, but from higher altitude and lower temporal resolution.

1.1 Motivation

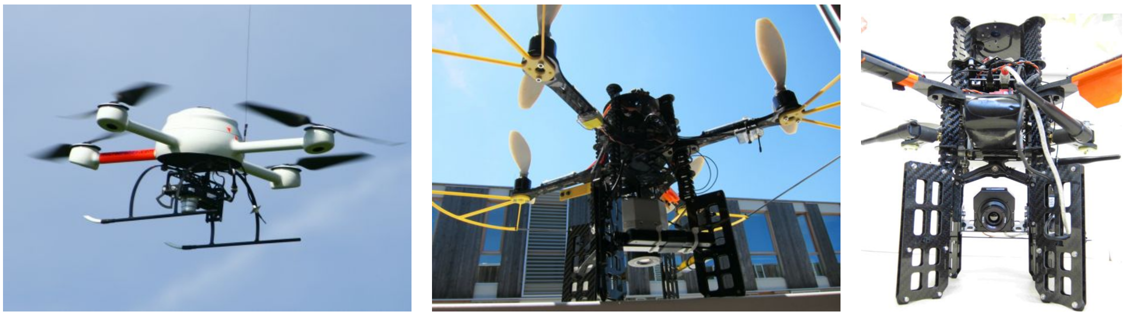
UAVs are widely used in the military domain. Advances in technology, material science, and control engineering made the development of small-scale UAVs possible and affordable. Such small-scale UAVs with a total weight of approximately 1 kg and a diameter of less than 1 m are getting prominent in civilian applications and pose new research questions. These UAVs are equipped with sensors such as accelerometers, gyroscopes, and barometers to stabilize the flight attitude and global positioning system (GPS) receivers to obtain position information. Additionally, UAVs can carry payloads such as visual and infrared (IR) cameras or other sensors. Figure 1.1 shows such UAVs with different sensors.

Thus, UAVs enable us to obtain a bird's eye view of an area which is helpful in many applications such as environmental monitoring, surveillance and law enforcement, border control, farmland and crop monitoring, object detection, construction sites assessment, and disaster management [45, 43]. In such scenarios we can not rely on a fixed infrastructure and therefore the available information (e.g., maps) may no longer be valid. The overall goal, hence, is to provide a quick and accurate overview of the affected area, typically spanning hundreds of thousands of square meters. This overview image is refined and updated over time and can be augmented with additional information such as detected objects or the trajectory of moving objects. When covering large areas at reasonable resolution from such small-scale UAVs, the overview image needs to be generated from dozens of individual images. Moreover, a number of UAVs equipped with cameras is employed instead of a single UAV to cope with the stringent time constraints and the limited flight time. The UAVs—flying at low altitudes of up to 100 m—provide images of the target area which are mosaicked to an accurate overview image. This process is refereed to as image mosaicking which is a noteworthy application of aerial imagery for further information retrieval from the target area.

In this thesis we describe different methods and their trade-offs for generating mosaics in order to surveil a certain area. We present different approaches which allow to quickly mosaick the individual images and refine the alignment over time as more images are available. Note that in sensitive cases of surveillance each image might have critical details which need to be retained even after the image is placed

[1] http://maps.google.com/

[2] http://www.bing.com/maps/



(a) MD4-200 with a visual camera.

(b) AscTec Pelican with an FLIR Photon 640 thermal camera.

(c) AscTec Pelican with an FLIR Tau-2 thermal camera.

Figure 1.1: Different UAVs used for acquiring thermal and visual images.

in a mosaic. In cases where UAVs are supposed to fly and take images without any loop in their route (e.g., border control, road construction and object following) the problem of mosaicking and orthorectification gets more challenging. In our methods we cope with the limitations of small-scale UAVs. We also consider the multispectral aerial imagery with heterogeneous sensors.

1.2 Problem statement

We aim to utilize small-scale UAVs to construct an overview image of a target area with different sensors (visual or thermal cameras). To construct this overview image we obtain a set of individual images based on a pre-planned mission. The accuracy of the constructed overview image by mosaicking the individual images is of significant importance. In most of the applications we need our mosaic to be orthorectified (relative distances are preserved) and georeferenced (the location in terms of coordinate systems is established). In this thesis we will explain how to achieve such a result considering the following limitations: limited payload, limited flight-time (since they are battery powered), limited flight altitude, and varying weather condition. These limitations force UAVs to just take images at individual predefined locations. This causes different angles of view looking to the same scene which intensifies the problem of mosaicking with non-planar objects.

Considering the mentioned scenario we address and solve the following three problems:

- How to exploit the metadata (data from inertial measurement unit (IMU) and GPS) in mosaicking of aerial image taken by small UAVs? The metadata per se are not sufficient for mosaicking. Therefore we explore the possibility of exploiting the metadata and the image processing to improve the mosaicking.

- How to produce orthorectified mosaics without using the metadata? We aim to discover the sources of errors in image mosaicking and mitigate them. In this approach we are not relying on image metadata or any global optimization method. Yet, we aim to produce orthorectified mosaics even with low overlap ratios and no loop(s) in the image sequences (cp. Section 2.3.3).

- How to perform a robust interspectral registration between images taken from heterogeneous sensors (e.g., thermal and visual cameras)? The images from such different types of sensors contain different characteristics and features. We aim to find the common features and correspondences between these images in order to register them together.

1.3 Contribution

The main contribution of this thesis is three-fold:

- First, we show how to increase the orthorectification by using a hybrid method of combining UAV metadata with image processing. Since the conventional mosaicking approaches are computationally expensive, we identify their bottlenecks and by exploiting the metadata such as GPS and IMU we reduce their complexity. In our hybrid approach we use the GPS data to find the adjacent images and their approximate positions in the final mosaic. We also approximate the orientation of the camera by using the IMU data. This helps us to rectify the images for final mosaicking. The contribution of this part has been achieved mutually with my colleague Daniel Wischounig-Strucl and the results are published in [79] and are also issued in a patent application [49].

- Second, we focus on image processing without considering the metadata. We construct an incremental mosaic while preserving the orthorectification and accuracy as much as possible. Most of the existing works handle the accumulated error by global optimization (which tries to distribute the accumulated error), local optimization, and blending algorithms to make the mosaic visually appealing. Instead, we study the sources of error, and we show how to minimize these accumulated errors. This work is published in [77]. Exploiting a reference plane to mitigate the mosaicking error is issued in a patent application [48].

- The final contribution of these thesis concentrate on the registration of thermal and visual images, which includes:
  - proposition of a general method for lens distortion correction of the thermal cameras (published in [76]),
  - introduction of feature descriptors for robustly identifying correspondences between images of different spectrums,

   - registration of image mosaics,
   - registration based on depth maps.

  In more detail, we introduce a robust feature along the edge and demonstrate the improvement of identifying corresponding points based on this feature in the general case. We further propose two methods to improve the registration of low-altitude aerial images. The first method exploits visual and thermal image mosaics, whereas the second method exploits the depth map of the scene to perform feature extraction and registration.

We have implemented and tested all these approaches in our UAV system. Eventually, we perform a quantitative evaluation over different image registration or mosaicking methods. Additional minor contributions within the scope of our project is published in [47, 46].

1.4 Thesis outline

The remainder of this thesis is organized as follows. Chapter 2 provides the prerequisites to follow the rest of the thesis. This background explains the camera architecture, image acquisition techniques, aerial imagery, image mosaicking techniques and our multi-UAV project. In Chapter 3 the state of the art in the fields of aerial image mosaicking and interspectral image registration and their differences to our work are presented. In Chapter 4 we propose two methods for mosaicking aerial images from small-scale UAVs. This chapter explains how to promptly generate orthorectified mosaics, with or without metadata. Chapter 5 focus on interspectral registration and performing multispectral mosaicking of aerial images. In other words, aerial images from different spectrums (i.e., visual and thermal images) are registered together.

The results and discussion over the mentioned topics are presented in Chapter 6. It includes the evaluations of our methods, sample results of our mosaicking and registration methods, quantitative metrics, and expansion of the discussion. The thesis is concluded by Chapter 7 which summarizes the goals, contributions, results, and the future work.

CHAPTER 2

Background

In this thesis we assume that the readers have basic knowledge of image acquisition and image processing. However, we summarize some concepts which will be used in the context of the thesis. This chapter explains the concepts of optics and image acquisition in a nutshell and describes the basics of image mosaicking in more details.

2.1 Cameras and digital imaging

In general, the electromagnetic waves emitted from an object make that object visible to the sensor which is made for that type of radiation. Figure 2.1 shows variations of different electromagnetic radiations. In the scope of our work we only use images acquired from visible light or IR radiation.

The early images were produced by letting the light emitted from a scene to pass through a hole (aperture) into a light-proof box or chamber. The rectilinear propagation of the light creates the image of the scene on the opposite side of the box or chamber. This is known as the basic description of a *pinhole camera*, also known as camera obscura. Later it was perfected by a converging lens and the capability of image recording, which led to what we know as conventional camera.

Cameras are convenient sensors which are made to capture and record a specific range of electromagnetic radiation. For instance, a visual camera (also referred to as RGB) records the visible spectrum shown in Figure 2.1. The output of such recording is an images which represents a specific scene captured prior in time. Image acquisition process is susceptible to errors similar to other measuring and representing methods. Advances of technology introduced a numeric representation of an image known as digital image. Since we have finite resources we represent a digital image in a discrete way. This is done by using matrix $\mathbf{I}$ comprising the intensity values in each element as image pixel. This matrix is two-dimensional for grayscale images and three-dimensional for color images. In the form of pixel representation, $I(x, y)$ represents a grayscale digital image, while $I(x, y, b)$ represents a color image, where parameter $b \in \{1, 2, 3\}$ implies the red, green or blue band of

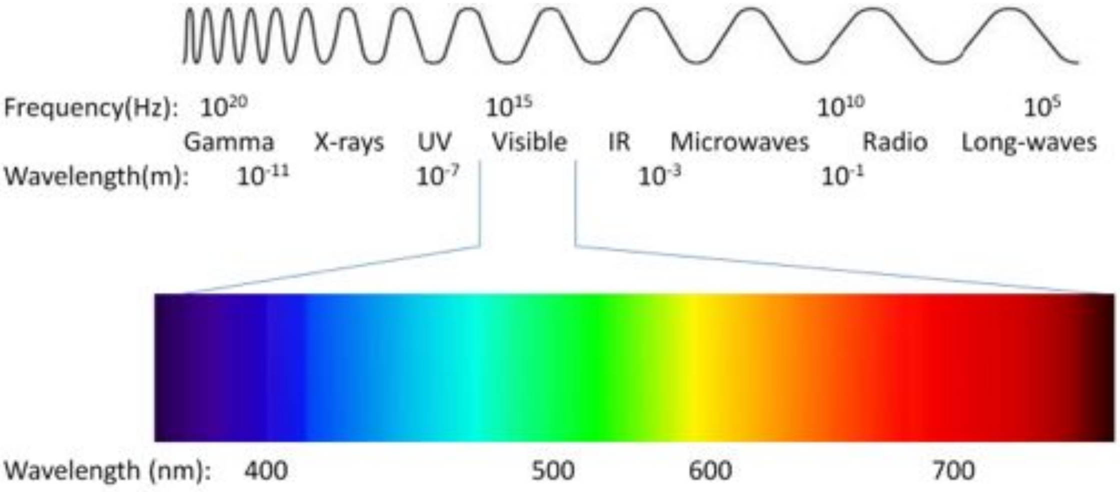


Figure 2.1: Electromagnetic spectrum.

the visible light. More detailed descriptions of this topic can be found in digital imaging textbooks such as [7, 68].

2.1.1 Infrared imaging

Visual images are the most popular type of images, because they represent the true colors seen by the human eye. Nonetheless, IR images have their own advantages. Since IR radiation has a longer wavelength in comparison with visible light (cp. Figure 2.1), it can penetrate better in fog or smoke. Furthermore, the IR radiation from heat makes it possible to sense human or any object with higher temperature than environment, even at night. Infrared radiation range starts with the wavelength of $0.7\,\mu m$ and extends up to $1\,mm$, which is divided into three different categories:

- Near infrared (NIR), $0.78 - 3\,\mu m$
- Mid infrared (MIR), $3 - 50\,\mu m$
- Far infrared (FIR), $50 - 1000\,\mu m$

Aside from visual cameras in this thesis we utilize thermal cameras as our aerial sensors, which operate in MIR. Two samples of small FLIR [1] thermal cameras are shown in Figures 1.1(b) and 1.1(c). Thermal cameras on average have far less resolution as compared to visual cameras. More detailed descriptions of this topic can be found in infrared imaging textbooks such as [17, 70].

[1] http://www.flir.com/

2.1.2 Lens distortion

Unfortunately, all camera lenses produce a type of distortion in which straight lines appear curved. Cameras with wider lens angles or field of view (FOV) show higher radial distortion. Brown's distortion model [6] formulates the radial and tangential distortion including the principal point estimation. Let $\mathbf{P} = (x, y)$ be a normalized point in image reference system, the undistorted point $\mathbf{P_u}$, using a 6th order radial and 2nd order tangential model can be acquired by

$$\begin{aligned}\mathbf{P_u} = \begin{bmatrix} x_u \\ y_u \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + (k_1 r^2 + k_2 r^4 + k_3 r^6) \begin{bmatrix} x_n \\ y_n \end{bmatrix} + \\ \begin{bmatrix} 2k_4 x_n y_n + k_5(r^2 + 2x_n^2) \\ 2k_5 x_n y_n + k_4(r^2 + 2y_n^2) \end{bmatrix},\end{aligned} \tag{2.1}$$

where point (x_n, y_n) is in a coordinate system considering the principal point, $\mathbf{PP} = (PP_x, PP_y)$, as its origin $(x_n = x - PP_x, y_n = y - PP_y)$, $r = \sqrt{x_n^2 + y_n^2}$ represents the distance from principal point, k_1, k_2 and k_3 are the radial distortion coefficients and k_4 and k_5 are the tangential distortion coefficients.

2.2 Remote sensing

Remote sensing is the process of obtaining information about objects or phenomena from a remote location without any physical contact. In modern usage, the term is used in the concept of sensing the electromagnetic radiation from the objects on earth through the sensors mounted on airborne or spaceborne platforms. Although satellite imagery has covered a large portion of remote sensing, utilizing small-scale UAVs for remote sensing is growing rapidly. UAVs, commonly known as drones, are aircrafts without human pilot on board. Instead, they are controlled either remotely by human pilots or autonomously by predefined planned missions. Based on their structure, size and capabilities, they are classified to different groups. They can be fixed-wing like airplanes or rotary-wing which can maneuver easier. Remote sensing by UAV is cheaper and easier to deploy. It can facilitate many applications of remote sensing such as surveillance, agricultural monitoring, construction monitoring, and disaster management.

2.3 Image mosaicking

An image mosaic is an image which is built from a set of smaller individual images. Image mosaicking consist of the steps necessary for constructing such mosaics and includes:

1. **Finding adjacent (neighboring) images.** Figure 2.2 shows a sample of 3 images to be mosaicked and their adjacency graph. Each image represent a

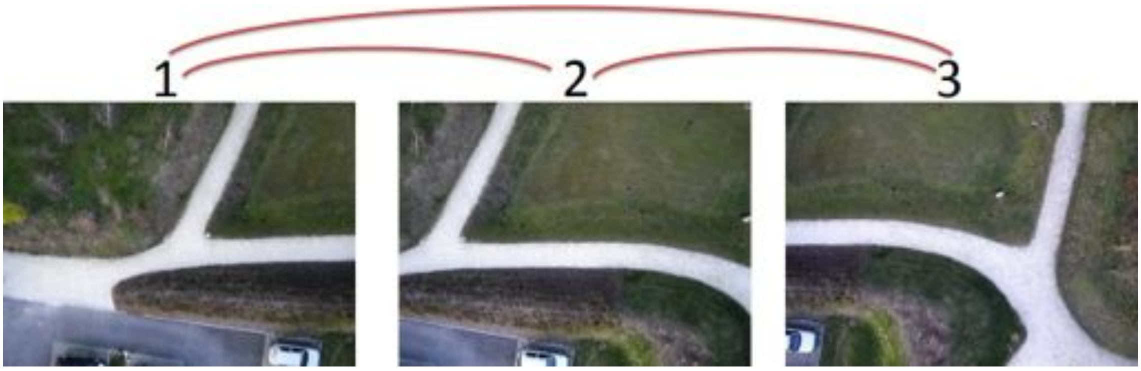


Figure 2.2: Sample three images to be mosaicked. In adjacency graph shown in red, images are the vertices and adjacent images are connected with an edge.

vertex (node) of the graph and the vertices are connected by an edge if the corresponding images are adjacent (have overlap). The adjacency of two image can be determined in different ways such as using image processing tools or using the metadata. For instance, two images can be tagged as adjacent if their GPS data show a relative distance less than a threshold.

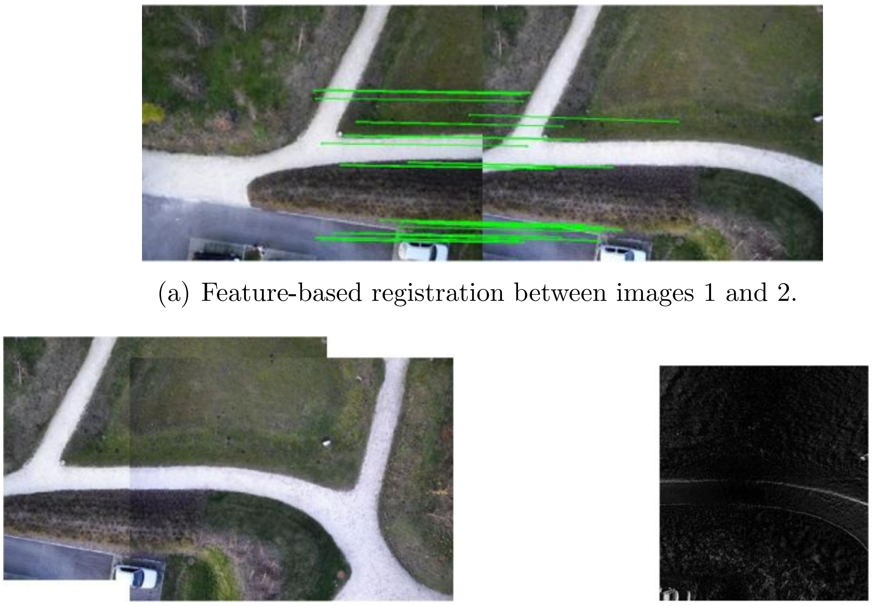
2. **Finding the correspondences between images and registration.** The first step for registering images together is to find their correspondences. It can be done by feature-based methods as shown in Figure 2.3(a) or by direct (pixel-based) methods as shown in Figure 2.3(b). In Figure 2.3(c) the differences of the intensity values between two images in the overlapping area is visualized. The brighter the color, the higher the deviation of the registration at those areas is.

3. **Aligning those images.** In this step the transformations are calculated. The transformations (e.g., translation, similarity, affine) are used to transform all images into one coordinate system and align them. Figure 2.4(a) shows a sample of image alignment.

4. **Stitching them together.** The final step is to stitch the aligned images and construct a seamless mosaic. Figure 2.4(b) depicts such mosaic.

The direct method of registration considers all pixels and measures the correlation between two image pixels. Although this method can be optimized (e.g., by exploiting image pyramids), it usually performs slow. In addition, it is not robust against scale, rotation, and in general affine transformation.

On the other hand, the feature-based alignment methods are faster and more robust. Considering images I_n, I_m, feature-based and pairwise (considering only one pair at a time) mosaicking is typically performed with the following steps:

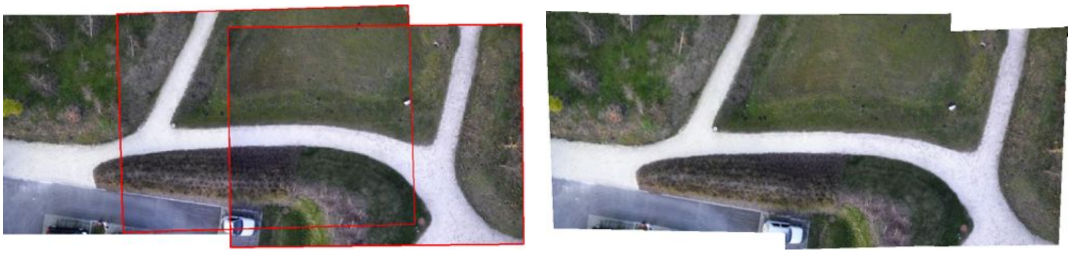


(a) Feature-based registration between images 1 and 2.

(b) Pixel-based registration between images 2 and 3. Different possible alignments are checked until the maximum correlation is achieved.

(c) Visualizing the pixel intensity differences in the overlapping area between images 2 and 3.

Figure 2.3: Finding the correspondences between images and registering them together.



(a) Aligning the images.

(b) Stitching all images together.

Figure 2.4: Final mosaicking steps.

- **Correcting the internal geometric distortion.** Brown's distortion model [6] can tackle the radial and tangential distortion including the principal point estimation.

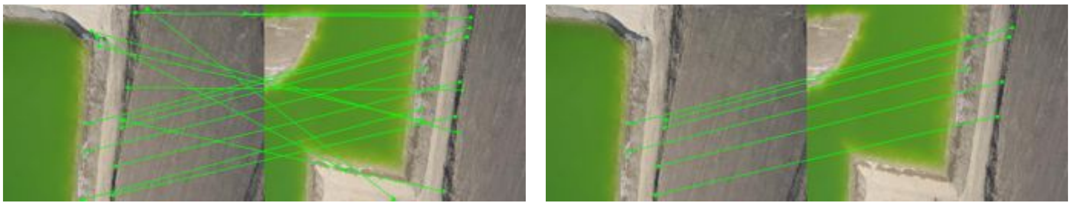
- **Feature extraction and image registration.** Different methods can be used to extract features which is later used for image registration (e.g., by using scale-invariant feature transform (SIFT) [36], speeded up robust features (SURF), or Harris corner[20]). Features extracted from the new unregistered image are matched with the previously registered image. Traditionally correspondences are determined by computing the similarity between descriptor vectors associated to each point. Figures 2.5 to 2.8 show some samples of feature matching by SIFT and SURF methods. As shown in these figures, different parameters and thresholds affect the position and total number of the features. Assume $\tilde{\mathbf{x}}_n$ and $\tilde{\mathbf{x}}_m$ are sample feature points (in homogeneous coordinates) respectively in images I_n and I_m. The feature matching function $R(\tilde{\mathbf{x}}_n, \tilde{\mathbf{x}}_m)$ shows the correspondences (matching pairs), which in this thesis we refer to it as the registration function.

- **Defining the projection model.** A *projection model* defines how to project an image of a three-dimensional scene onto a planar surface. Projection models are explained in Section 2.3.2. In the scope of our work for simplicity we mainly assume a planar model, since we mostly fly over areas with a dominating ground plane. Clearly, there might be non-planar objects on the ground which we will discuss in Section 4.2.2.

- **Defining the appropriate transformation.** Based on different scenarios we can choose between different existing transformations such as translation, similarity, or projective transformation (homography). The transformation (in homogeneous coordinates) from the coordinates of image I_n to the coordinates of image I_m is defined by $\tilde{\mathbf{H}}_{I_n,I_m}$. $\tilde{\mathbf{H}}$ is a 3×3 matrix which represents the relative rotation, scale, translation and projection. By definition we have $\tilde{\mathbf{H}}_{I_n,I_m} = \tilde{\mathbf{H}}^{-1}_{I_m,I_n}$.

- **Removing outliers and calculating the transformation function.** In statistics, *outliers* (as opposed to *inliers*) are observations that are markedly deviated from the rest of the data. In our research, an outlier is a feature point with a deviation more than a threshold ϵ from its matched feature point after transformation. RANdom SAmple Consensus (RANSAC) and least median of squares (LMS) are usually used to remove the outliers and find the best fitting transformation parameters. Each iteration of RANSAC estimates a transformation while trying to find the maximum set of inliers as a subset of all matched pair-points while the equation

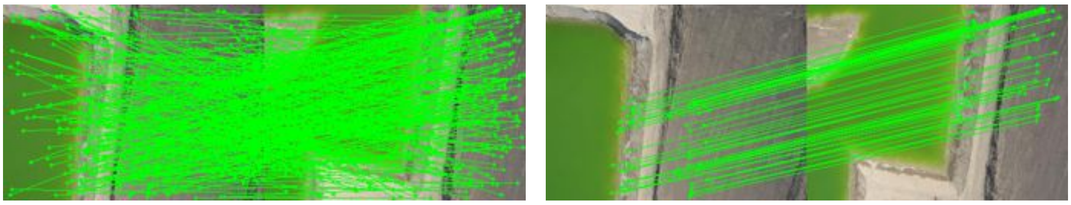
$$\|\mathbf{d}_i\| = \|\tilde{\mathbf{H}}_{I_n,I_m}\tilde{\mathbf{x}}_{m_i} - \tilde{\mathbf{x}}_{n_i}\| \leq \epsilon \tag{2.2}$$



(a) Total number of matches = 35.

(b) Number of inliers = 7.

Figure 2.5: Sample of SIFT feature matching with a peak-selection threshold 0.02.



(a) Total number of matches = 542.

(b) Number of inliers = 55.

Figure 2.6: Sample of SIFT feature matching with a peak-selection threshold 0.0001.

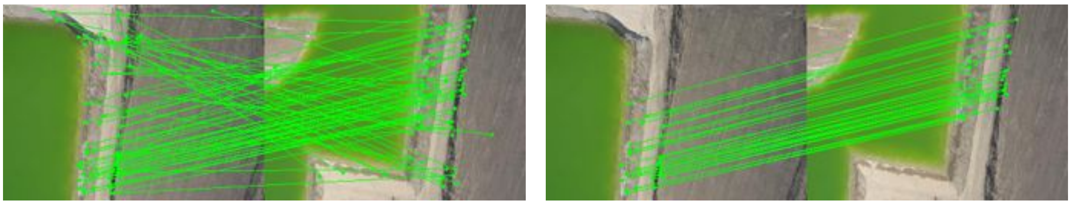
holds true, where i is the index of the inlier, ϵ is a threshold which varies based on application, $\mathbf{d}$ shows the disparity vector between points $\tilde{\mathbf{x}}_{n_i}$ and their estimated corresponding position and $\tilde{\mathbf{H}}$ is the optimized transformation matrix to change the coordinates between two images. The iteration is repeated for a predetermined number of cycles and the final solution is the sample homography and set of the points with the largest number of inliers.

- **Transformation and alignment.** In this step we transform the new image to the coordinates of previously registered image and perform the resampling (by using interpolation). Note that sometimes image registration is also considered as a part of this step, because in some sources, image registration is not only mapping the correspondences but also transforming all data into one coordinate system.

- **Mosaic construction.** Finally, in order to build the incremental mosaic, we merge the transformed image with the mosaic constructed so far.

2.3.1 Panorama vs. mosaic

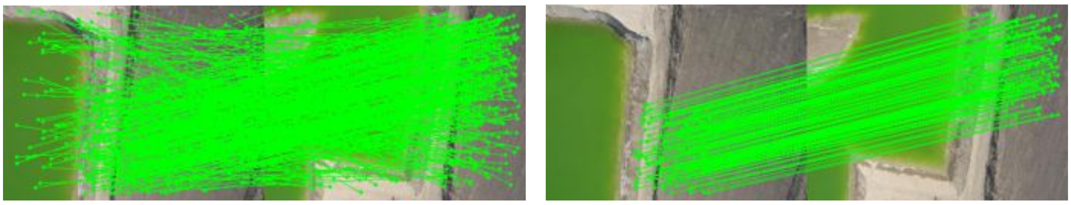
Although the terms panorama and mosaic are used interchangeably, we define them in different concepts. Panorama is an extension of FOV while mosaic is an extension to point of view (POV). Figures 2.10 and 2.9 depict this concept.



(a) Total number of matches = 105.

(b) Number of inliers = 34.

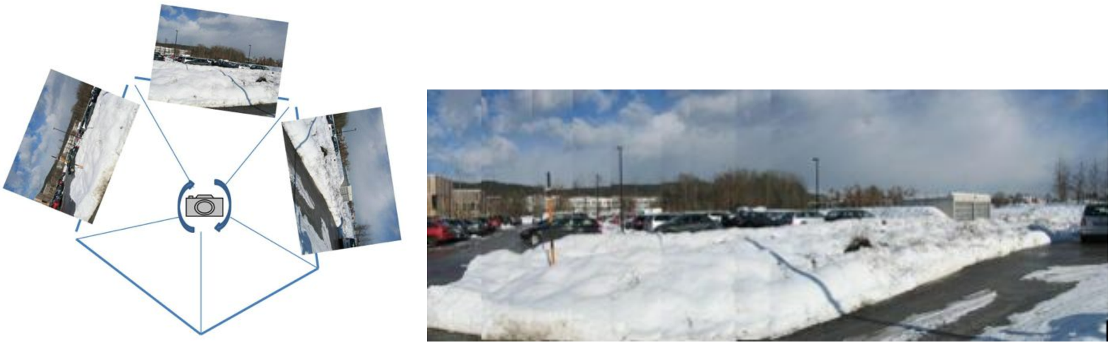
Figure 2.7: Sample of SURF feature matching with a blob-response threshold 1.



(a) Total number of matches = 582.

(b) Number of inliers = 134.

Figure 2.8: Sample of SURF feature matching with a blob-response threshold 0.001.



(a) Panorama is extension of FOV. Camera rotates around its optical center.

(b) A sample panoramic image.

Figure 2.9: Structure of image panorama.

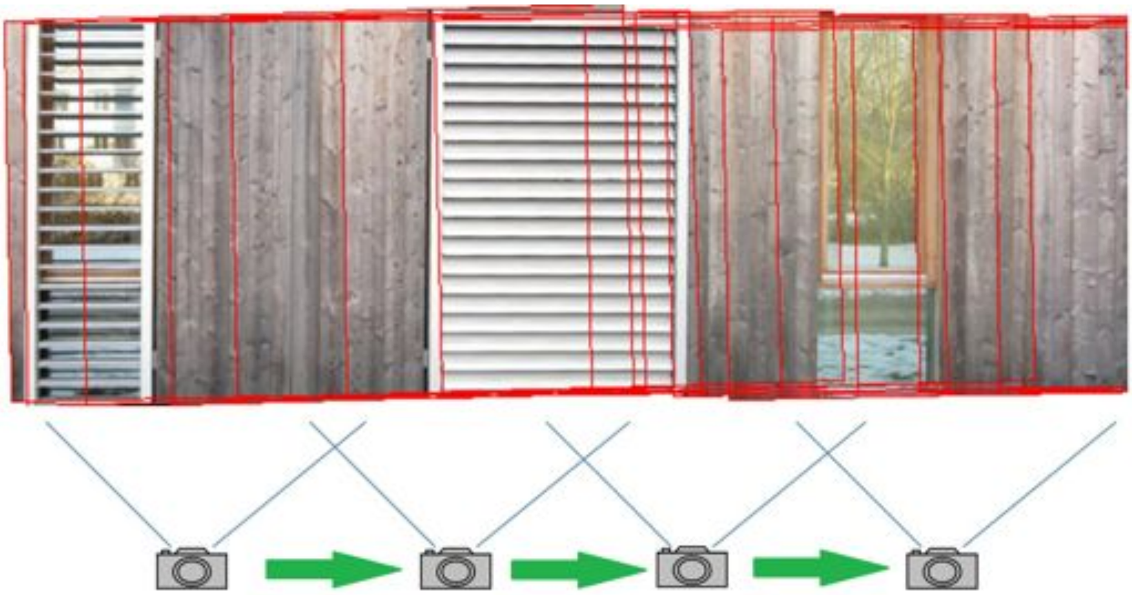


Figure 2.10: Structure of image mosaic as an extension to POV.

Most of panoramic images are taken in a way that the camera rotates around its optical center. Since the camera does not change the location, all the images are taken from the same point of view and it mitigates the problem of parallax. However panoramic construction has it own challenges to find the appropriate projection model based on the type of rotation of the camera, type of the scene and the intended shape to warp the panorama to. On the other hand, mosaics are built by aligning and stitching the images which are taken from different points of view (e.g., aerial images taken by moving UAVs). Figure 2.4 also depicts a mosaic, since the aerial images are taken from different UAV picture-points.

2.3.2 Projection models

The projection models define how to project an image of a three-dimensional scene onto a two-dimensional (planar) surface. Converting different coordinate systems needs an appropriate projection model based on application. In image mosaicking we have usually a dominant ground plane of the earth and therefore we can use a planar projection. Though, the most appropriate projection for aerial imagery would be the parallel projection. Because, in this type of projection we have the nadir-view for each individual object or pixel. In other words a parallel projection is equivalent to a perspective projection with an infinite focal length. To get as close as possible to a parallel projection, in real world, either we need a quite planar ground or we need more number of picture-points from nadir-view.

Spherical projection is mainly used in cases such as panoramic image construction when the camera rotates around its optical center. Warping the images into spherical coordinate models appropriately this type of image acquisition. When in a panoramic image the camera is rotated around it axis, the cylindrical projection is the most appropriate type.

2.3.3 Loop in image sequence

As explained in Section 2.3, we assign an adjacency graph to a set of images, where images are the vertices and images with overlap (adjacent images) are connected with an edge. Now we define the term *loop* in the image sequence. We say there is a loop in a set of images if there is a simple cycle in its adjacency graph. Likewise we say there is no loop in the image sequence if its adjacency graph is a forest. For instance the three images shown in Figure 2.2 have a short loop, since their adjacency graph has a loop of length 3. Images with high overlap ratio usually have such short loops. If the first image and the third image did not have any overlap (were not adjacent), then it would be a loop-free sequence. In Chapter 6, we show a sequence of images with a loop of length 37 (Figure 6.3).

2.3.4 Optimization methods and seamless mosaicking

Given a set of images with their inlier points in the pairwise overlapping area, *global alignment* (global optimization) is the process of finding the appropriate parameters and homographies which minimize the registration and disparity errors between all pairs of images. The term is generalized to *bundle adjustment* [67] when the process is done in three dimension. When the number of images (n) increases, the number of distinct pairs ($n(n-1)/2$) grows quadratically. Thus, minimizing the error by considering all pairs of images becomes more complex. Heuristic methods are alternatives to simplify the optimization process. The global optimization methods are useful when there is a loop in the image sequence. Without loop(s), the mosaicking process reduces to a simple pairwise mosaicking and the global optimization does not improve the image mosaicking at all.

Once the global optimization is done, the local optimization methods are performed to reduce the local displacement and registration errors. Parallax removal and deghosting are samples of local optimization. Image gain correction and extrapolation are two samples of other methods for improving the visual quality of the images.

2.4 Multi-UAV project

This work was performed as part of the project *Collaborative Microdrones (cDrones)*[2]. The basic idea of the project is to deploy multiple small-scale UAVs to support first

[2] http://uav.aau.at/

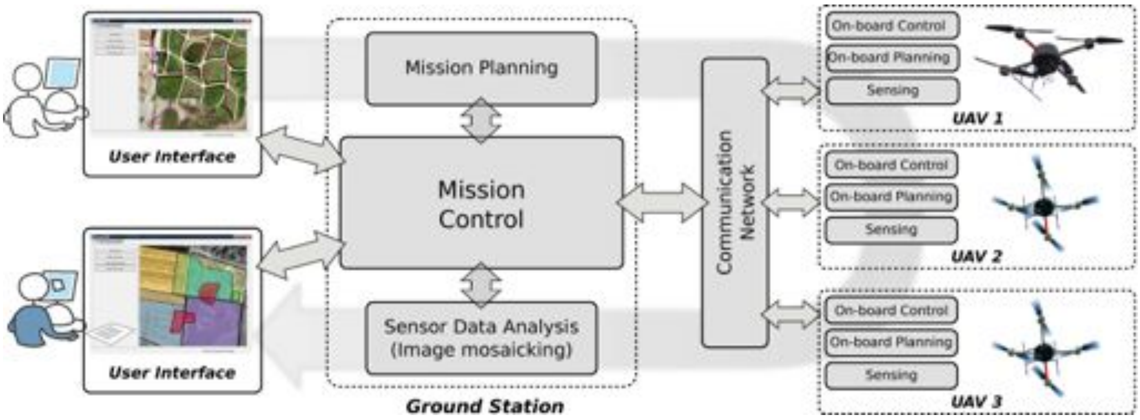


Figure 2.11: System architecture of our multi-UAV project.

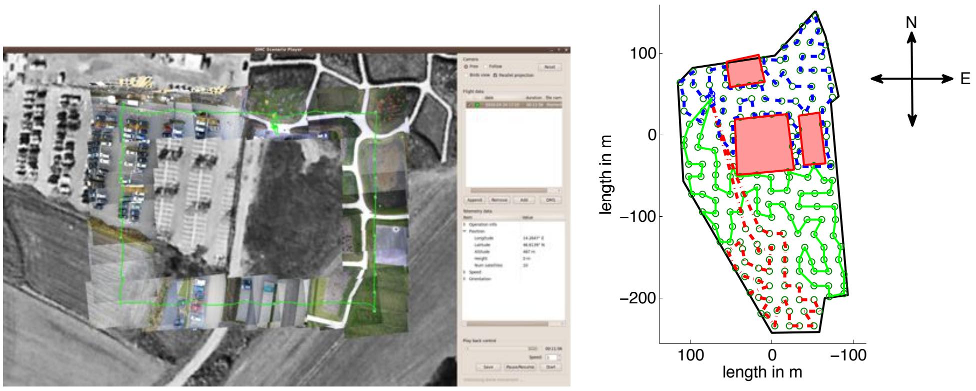
responders in disaster assessment and disaster management. In particular we use commercially available quadrocopters since they are agile, easy to fly and stable in the air due to their on-board sensors such as GPS and IMU. Each UAV is equipped with a camera. Figure 2.11 describes the system architecture of the project.

The intended application can be sketched as follows: The operator first specifies the areas to be observed on a digital map and defines the quality parameters for each area [45]. Quality parameters include the spatial and temporal resolution of the generated overview image, and the minimum and maximum flight altitude, among others [43].

Based on the user's input, the system generates plans for the individual UAVs to cover the observation areas [44]. Therefore, the observation areas are partitioned into smaller areas covered by a single picture taken from a UAV flying at a certain height. The partitioning has to consider a certain overlap of neighboring images which is required by the stitching process. Given a partitioning we can discretize the continuous areas to be covered to a set of so-called picture-points. The picture-points are placed in the center of each partition at the chosen height. The pictures are taken with the camera pointing downwards (nadir-view).

The mission planner component generates routes for individual UAVs such that each picture-point is visited taking into account the UAV's resource limitations. The images together with metadata (i.e., the position and orientation of the camera) are transferred to the base-station during flight where the individual images are stitched to an overview image. Figure 2.12 illustrates samples of mission planing and aerial image acquisition within the scope of the project. The restricted areas (e.g., buildings and dangerous areas) are marked as obstacles. After planning is finished, the mission is executed. The UAVs take off fully autonomously, fly the specified routes and send the pictures to the ground station.

Mosaicking the individual images and constructing an overview image is the final step of the project. In our research, we are not considering any type of global or



(a) GUI showing new images over outdated background.

(b) Sample mission showing restricted areas and planned routes.

Figure 2.12: User interface and mission planing.

local optimization for image mosaicking. The main reason is to save the computational power and present a real-time incremental mosaic. Another reason is that the global optimization needs enough pairwise overlap and presence of loop in the image sequence, while these prerequisites may not hold in our scenarios. Local optimization and other methods which aim to produce a visually appealing mosaic are not also appreciated. Because these procedures sometime remove some information which may be important for surveillance, detection or monitoring purposes.

Additionally, various applications within the scope of the project are introduced. These applications include object detection and tracking, multi-UAV area coverage to help the first responders for disaster management, construction site monitoring, and advertising. Sample images, videos and demonstrations are available on the project web-page [3].

[3] http://uav.aau.at/

CHAPTER 3

State of the art

It is important to understand the challenges of aerial image mosaicking which are mostly determined by the type of the imagery. Important aspects to consider include the following questions:

- Are the images taken from satellites or at lower altitudes?
- What type of sensors are used and what is the difference of the wavelengths between different electromagnetic bands?
- What is the noise level and resolution of the achieved images?
- Is image metadata available?
- Are the images taken at the same time?
- Are the images taken from same point of view?
- How large is the overlap between images?
- What is the dominant transformation between images (relative translation, scale, rotation and perspective)?
- Is the scene flat and how to project a three-dimensional scene to an image?
- What are the quality of service (QoS) requirements (resolution, orthorectification, georeferencing, state of visual appealing and being seamless, etc.)?

By considering the mentioned aspects, we present the related works in different fields of aerial images mosaicking and interspectral image registration.

3.1 Related work in aerial image mosaicking

Much research has been done in the area of mosaicking of aerial imagery and surveillance over the past years. Many approaches have been proposed ranging from using low-altitude imagery of stationary cameras and UAVs to higher altitudes imagery captured from balloons, airplanes, and satellites. High-altitude imagery and on-ground mosaicking such as panoramic image construction are not in our area of interest since they deal with different challenges. In this thesis we are focused on aerial images obtained from low-altitude UAVs. A huge number of aerial image mosaicking approaches rely on medium to large UAVs. These UAVs have more capabilities in aspects of their computational power, data transmission rate, payload capacity, accuracy of measurement devices, and flight time. Based on these parameters a variety of approaches are proposed for mosaicking of images taken from UAVs.

There has been a breakthrough regarding the seamless mosaicking in past years by exploiting robust feature extraction methods [82, 61, 3], depth maps [27, 8], 3D reconstruction of the scene, image fusion, and many other approaches (e.g., [63, 57]). However, few researchers have compared different mosaicking methods (e.g., [2]). Figure 3.1 shows a sample stitching of five sequential images generated by a SURF feature-based algorithm [3]. In this mosaic the images are aligned well but the obvious drawback is that the transformation performed on the images leads to a distortion in scales and relative distances. Such a traditional feature-based approach is difficult for our case because the generation of an orthorectified and georeferenced image is hardly possible due to the scale and angle distortions as well as the error propagation over multiple images. The non-planar surface is one of the main reasons for this distortion, i.e., by using corresponding points at different elevation levels for the image registration. In principle, it is possible to improve the mosaicking result by using metadata, global alignment and bundle adjustment [57, 15, 58], but we need to know either accurate IMU data of the UAV's camera or accurate corresponding feature pairs.

A challenge of low-altitude imagery and mosaicking for surveillance purposes is finding an appropriate balance between seamless stitching and georeferencing under consideration of processing time and other resources. Many approaches have been proposed to tackle these problems. Examples include the wavelet-based stitching [80], image registering in binary domains [18], automatic mosaicking by 3D reconstruction and epipolar geometry [35], exploiting known ground reference points for distortion correction [42], IMU-based multispectral image correction [25], combining GPS, IMU and video sensors for distortion correction and georeferencing [5] and perspective correction by projective transformation [81]. Some of these approaches differ from ours in a sense that they consider higher altitude [5, 18, 35, 42, 81], while others use different types of UAVs such as small fixed-wing aircrafts [32, 25, 23]. These aircrafts show less georeferencing accuracy caused by higher speed and degree of tilting (higher amount of roll and pitch). Zhu et al. [84] performed an

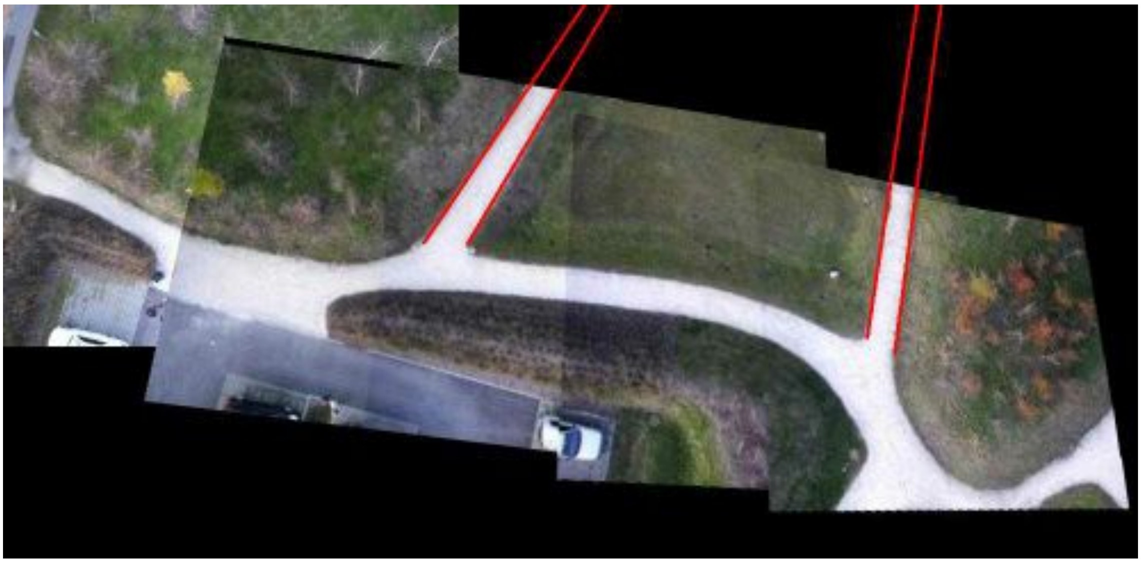


Figure 3.1: Mosaicking five images using SURF features. The path borders (red lines) are supposed to be almost parallel. This type of error accumulates over multiple images if not compensated.

aerial imagery mosaicking without any 3D reconstruction or complex global registration. The difference of their approach is that they used the video stream which was taken from an airplane. Huang et al. [23] performed also a seamless feature-based mosaicking using a small fixed-wing UAV, but no georeferencing assessment was conducted. Roßmann and Rast [50] also used small-scale quadrocopters. Their mosaicking results are seamless but lacking georeferencing. No details about the mosaicking approach are presented.

Schultz et al. [55] use a digital elevation model to mosaic images taken from an airplane. Hruska et al. [22] introduce an appropriate platform for small UAVs to be able to provide high resolution and georeferenced images by exploiting GPS and IMU. Afterwards they perform change detection by comparing different temporal images of a target area. In their work they remark the importance of internal geometric distortion correction but do not explain how it is used in mosaicking. Zhou [83] uses the video stream from a UAV (weight 10 kg) equipped with differential GPS, with an error range of a few centimeters, and real-time transmitter of video for further mosaicking purposes on the base-station. Xiang and Tian [75] also mention the role of high precision internal geometric distortion correction in georeferenced mosaic construction in addition to exploiting GPS and IMU. Agarwala et al. [1] cope with the problem of producing multi-viewpoint panoramas of long, roughly planar scenes but on the ground (e.g. the facades of buildings along a city street). They use Markov Random Field optimization to construct a composite from arbitrarily shaped regions of the source images, rather than building the panorama from strips of the source images. They also consider a higher pairwise overlap (with approximately 1 m distance between two picture-points) and the dominant plane of the photographed scene is defined by the user input.

3.2 Related work in interspectral image registration

Registering images acquired from heterogeneous sensors is not as simple as homogeneous image registration. Consider that two sensors record different ranges of electromagnetic spectrum from the same scene at the same time. Usually with larger wavelength distance between these two electromagnetic bands, the similarity between images decreases. In this case there is less mutual information and the probability of failure by conventional registration methods (cp. Section 2.3) increases. There is no ultimate solution for the problem of interspectral registration, since images from different spectrums may completely show different characteristics and features. However, many successful interspectral registrations have been performed by narrowing down the scope of the work (e.g., exploiting some prior knowledge about the images).

A few researches have focused on jointly registering thermal and visual images for the purpose of disaster site reconnaissance [52, 51]. Many early interspectral registration techniques have been employed to register the different bands of satellite images. Fonseca and Costa [14] propose an automatic satellite image registration based on wavelets. They mainly focus on the registration of images taken from the same sensor. Mahdi and Farag [37] propose a cooperative parallel optimization based on a genetic algorithm to match the different bands of multispectral satellite images. For time efficiency, their method demands a parallel implementation of the genetic algorithms and a supervisor process. Hong and Zhang [21] describe an automated registration technique by combining the feature-based and area-based matching for high resolution satellite images. They employ wavelet-based feature extraction and a relaxation-based image matching technique to reduce the local distortions caused by terrain relief. Although they manage to speed up the registration, they only consider the registration of panchromatic with multispectral images which are both almost in the same spectral range. H. Kim and M. Kim [29] focus mainly on the problem of parallax removal caused by different viewpoints. They improve the registration by correcting the terrain relief using a rigorous sensor model with precise sensor parameters and ellipsoidal height information extracted from digital elevation model (DEM). Kern and Pattichis [28] propose a robust interspectral registration using mutual-information models. The shape of the mutual-information surface is related to the frequency-domain characteristics of the imagery. Therefore, this mutual-information is used to iteratively optimize a target function and find the appropriate registration parameters. Lee [33] performs a coarse-to-fine multispectral satellite image registration. He uses the SURF method for fast initial feature extraction and handling the possible differences in scale and orientation. He then applies the Harris operator to extract more features. However, Teke [65] suggests to use the SURF feature extraction method to perform the whole process of the registration. Teke's method is simply taking advantage of the SURF parameters in cases when there are no rotation (Upright-SURF) or no scale differences (Scale-restricted SURF) between two images.

Satellite remote sensing is not the only field where interspectral registration plays a critical role. Medical imaging, object or face detection, surveillance and UAV remote sensing are other fields which are rapidly growing and fusion of different sensors become very handy. Schaefer et al. [53] perform multi-modal (thermal and visual) medical image registration and overlay. By exploiting prior knowledge of the human body they segment both images to find the matching body area. The authors of [31, 69] also exploit prior knowledge about the human face to register and fuse the thermal and visual face images. Istenic et al. [24] register the thermal and visual images of the facades of the buildings. Since most of the buildings are comprised of straight lines, they perform a hough transform over the images to extract those lines as the mutual features. Likewise, Coiras et al. [9] and Segvic [71] rely on having sufficient straight lines or structured polygons as a prerequisite for the registration. Du et al. [10] present an algorithm for automatic registration of the NIR and visual image sequences taken by a UAV. Since both cameras are mounted on a single UAV (i.e., the relative orientation is fixed), the computation of the extrinsic parameters of the cameras prior to flight is less expensive. Unlike all the mentioned works which are based on individual images for registration, Joo et al. [26] perform the registration by using sequence of the frames with moving objects. They first extract the moving region of each image as the target area, then they perform the matching and registration over that region. Despite the advantage of the method in some aspects, obviously the method fails in absence of the moving objects.

3.3 Differences to state of the art

In this section, we accent the differences between our work and state of the art by answering the questions from the beginning of this chapter.

In our work we focus on images taken from low-altitude (below 100 m) UAVs. As mentioned in Sections 1.2 and 2.4, we consider limitations such as inaccurate position and orientation information, non-nadir-view (tilt of camera), and limited resources. These limitations pose new challenges compared to other remote sensing scenarios such as satellite and airplane imagery. In high-altitude imagery, more sophisticated and accurate sensors (such as IMU, GPS, and laser scanners) are being used. In our work, we do not merely rely on GPS and IMU data, because these metadata are typically unreliable for small-scale UAVs. However, we consider both mosaicking methods either with or without exploiting the metadata.

Our methods do not rely on any prior knowledge about the scene and are able to deal with images with completely different spatial resolution, temporal resolution, scale, orientation, and a low amount of overlap. The scene is not necessarily flat and may include non-planar objects. When flying with UAVs at a relatively low altitude, non-planar objects on the ground make the feature matching and image registration more difficult. Taking images at individual predefined picture-points, causes different angles of view looking to the same scene and this intensifies the problem of non-planar objects. Yet, many other works for the purpose of real-time monitoring of wide areas rely on a single point of view. Most of these works focus on object detection and tracking by using moving cameras around a fixed center. In these cases, algorithms for active tuning of intrinsic camera parameters [39] or algorithms for background extraction and robust estimation of the displacements between consecutive frames [38] are of high importance.

In our work, we use both visual and thermal cameras. By thermal we mean a mid-wave infrared (MWIR) camera which is more challenging compared to NIR cameras. The thermal images are analog which have higher noise ratio compared to digital visual images. The thermal cameras that we use have $640 \times 512\,px$ spatial resolution. The resolution of digital images vary between $300 \times 400\,px$ and $3000 \times 4000\,px$. The resource limitations bound the spatial resolution, although the higher resolution is beneficial to most of the mosaicking projects. In some scenarios, we have to send the images efficiently (first the lowest resolution and, if possible, enhance it later) over the limited wireless network [73]. The images are taken mainly at different points in time. The temporal resolution range from seconds to years. The image overlap ratio is in the whole range of $0 - 100\%$.

Unlike most of the mosaicking methods we do not aim to produce a visually appealing mosaic. For instance, deghosting method (cp. Section 2.3.4) removes a ghost object because of its local inconsistency (e.g., object moves or changes over time). This improves the visual quality, while the removed information may be useful for the goal of the project. Hence, in the scope of our work (e.g., disaster management, surveillance, and overview map construction), we preserve the image integrity in the final mosaic. Besides, the mosaics are orthorectified and georeferenced.

So far we have mentioned the differences to state of the art by emphasizing on underlying assumptions of aerial imagery. At this point we explain the differences in the approaches and the methods for mosaicking. In our research, we consider an incremental system of mosaicking in which the images are added over the time and the previously mosaicked images are not repositioned. That is another reason that a global optimization is useless, since we do not have access to all images from the beginning. We also explore the quality of a loop-independent mosaicking. In other words, we study how to perform mosaicking is situations without any loop in the image sequence. In case of multispectral mosaicking, we introduce a robust feature descriptor for the purpose of interspectral registration. By this descriptor we extract mutual scale- and rotation-invariant features between visual and thermal images. In addition, we exploit the 3D structure of a scene to extract features which can be used for interspectral registration.

CHAPTER 4

Orthorectified mosaicking of UAV images

Orthorectification is the process of removing the effects of image perspective and distortion for the purpose of creating a geometrically correct image. An orthorectified image has a uniform scale and the relative distances are preserved. It can be used to measure the distances and angles accurately. Orthorectified images are widely used in virtual maps such as Google Earth, Bing Maps, etc. In this section we introduce some approaches to construct orthorectified mosaics from images taken form small-scale UAVs. As mentioned earlier in Section 2.4, providing an orthorectified mosaic in real time is the main goal for most of the scenarios in our project. The global optimization methods (cp. Section 2.3.4) are computationally expensive and they are only useful in presence of loops in the image sequence (cp. Section 2.3.3). The nature of our project urges the images to be taken at individual picture-points which cause a lower overlap ratio between them. With a smaller overlap ratio between images the global optimization methods become less effective. Furthermore, in cases of monitoring in which the image sequence is without loop(s) (e.g., monitoring pipelines or straight roads), we may not obtain an orthorectified and georeferenced mosaic due to the accumulation of the errors. In such scenarios we first present a hybrid method which combines the metadata with image processing to mosaick the images incrementally. Second we find the sources of error in pairwise mosaicking process and mitigate them.

4.1 Incremental mosaicking with a hybrid approach

This approach combines metadata-based and image-based stitching methods in order to overcome the challenges of low-altitude and small-scale UAV deployment such as non-nadir-view, inaccurate sensor data, non-planar ground surfaces, and limited computing and communication resources. For the generation of the overview image we preserve georeferencing as much as possible, since this is an important requirement for many applications. Our mosaicking method has been implemented on our UAV system and evaluated based on a quality metric.

The idea is to first place the new images based on the cameras position and orientation information on the already generated overview image. In the next step, we use image-based methods to correct for inaccurate position and orientation information and at the same time improve the visual appearance.

4.1.1 Problem definition and challenges

The major goal is to generate an overview image O_n of the target area given a set of n consecutive images $\{I_i | i = 1 \ldots n\}$. The overview image can be iteratively constructed by

$$O_i = Merge(O_{i-1}, I'_i), \tag{4.1}$$

where O_0 is an empty background (e.g., a zero matrix), I'_i is the transformed image of I_i by homography $\tilde{\mathbf{H}}_{I_i,I'_i}$ and the *Merge* function combines the transformed image to the overview image.

This mosaicking can be described as an optimization problem, in which we need to find $\tilde{\mathbf{H}}_{I_i,I'_i}$ in a way that it maximizes our quality function $\lambda(O_i)$. This quality function, based on our system scenario, balances the visual appearance and the georeferencing accuracy. While in some applications it is more important to have a visually appealing overview image, other applications may require accurate georeferencing in the overview image. We use a quality function that is a combination of the correlation between overlapping images and relative distances in the generated overview image compared to the ground truth. This quality function is explained in Section 6.1.1 in details.

In the following we summarize the most important challenges for solving our problem using images from low-flying, small-scale UAVs:

Low-altitude and non-planar surface. When taking images from a low altitude the assumption of a planar surface is no longer true. Objects such as buildings, trees, and cars cause high perspective distortions in images. Without a common ground plane, the matching of overlapping images requires depth information. Image transformations exploiting correspondences of points at different elevations may result in severe matching errors.

Non-nadir-view. Due to their light weight small-scale UAVs are vulnerable to wind influences requiring high-dynamic control actions to achieve a stable flight behavior. Even if the on-board camera position is actively compensated, a perfect nadir-view of the images cannot be provided.

Inaccurate position and orientation information. The UAV's auxiliary sensors such as GPS, IMU, and altimeter are used to determine its position and orientation. However, such auxiliary sensors in small-scale UAVs provide only limited accuracy which is not comparable with larger aircrafts. As consequence, we can not rely on accurate and reliable position, orientation and altitude data of the UAV. Hence, we have to deal with this inaccuracy in the mosaicking process.

Resource limitations. In our application the resources such as computation power and memory on-board the UAVs and also on the ground station are very limited. In disaster situations it is usually not possible to have a huge computing infrastructure available. The base-station typically will consist of notebooks and standard desktop PCs. But at the same time, we want to present the overview image as quick as possible.

Incremental refinement. The individual images are taken from multiple UAVs in an arbitrary order. An incremental approach is needed to present the user the available image data as early as possible while the UAVs are still on their mission. The more images are taken the better the overview image gets. This also means that a new image may require to adjust the position of already processed images to improve the overall quality.

4.1.2 Mosaicking approach

As described in Section 4.1.1 we must find the appropriate transformation $\tilde{\mathbf{H}}_{I_i,I'_i}$ for each image I_i captured at a picture-point in order to solve our mosaicking problem. There are two basic approaches for computing these transformations: The *metadata-based approach* exploits auxiliary sensor information to derive the position and orientation of the camera which is then used to compute the transformations. In this case we assume that auxiliary sensor data (i.e., GPS, IMU, altitude, and time) is provided for each captured image. The *image-based approach* only exploits image data to compute the transformations. In this section we first present the basic approaches considering the challenges of small-scale UAVs and then describe our hybrid approach which enhances metadata-based alignment with image-based alignment. The presented approaches vary in their resource requirements and their achieved results. Thus, they fit nicely to our problem domain.

Position-based alignment

A very simple and naive approach is to align the images based on the camera's position. Hence, for image alignment the world coordinates of the camera are mapped to corresponding pixel coordinates in the generated overview image. Defining the origin of the overview image of the observed target area as $\mathbf{o}_{world} = (lat, lon, alt)^T$ in world coordinates, all image coordinates are related to this origin on the local tangent plane (LTP) by approximation to the earth model WGS84.

Given the camera's position we compute the area covered by the picture in world coordinates relative to the origin taking into account the camera's intrinsic parameters. The relative world coordinates are directly related to the pixel coordinates in the generated overview image. An example of the resulting overview image is depicted in Figure 4.1(a) utilizing the placement function given in Equation 4.1. The transformation $\tilde{\mathbf{H}}_{I_i,I'_i}$ is reduced to a simple translation (with two degrees of freedom) for each image, i.e.,

$$\tilde{\mathbf{H}}_{I_i,I'_i} = \begin{bmatrix} 1 & 0 & T_x \\ 0 & 1 & T_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{4.2}$$

where T_x and T_y show the translation components (displacement in x and y directions) in the coordinates of the overview image. In this approach we assume reasonably accurate position information and a nadir-view but do not take into account the camera's orientation. The scale differences introduced by altitude differences is not compensated with this approach.

Position- and orientation-based alignment

A more advanced approach is to extend the naive position-based alignment by compensating the camera's orientation deviation (i.e., roll, pitch, yaw angles). The placement function of the individual images to generate the overview image is the same as in Equation 4.1. But instead of considering only translation, we use a perspective transformation $\tilde{\mathbf{H}}_{I_i,I'_i}$ with eight degrees of freedom, i.e.,

$$\tilde{\mathbf{H}}_{I_i,I'_i} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix}. \tag{4.3}$$

If we assume a nadir-view (i.e., zero roll and pitch angles) the transformation $\tilde{\mathbf{H}}_{I_i,I'_i}$ is reduced to a similarity transformation with four degrees of freedom, i.e.,

$$\tilde{\mathbf{H}}_{I_i,I'_i} = \begin{bmatrix} s\cos(\theta) & -s\sin(\theta) & T_x \\ s\sin(\theta) & s\cos(\theta) & T_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{4.4}$$

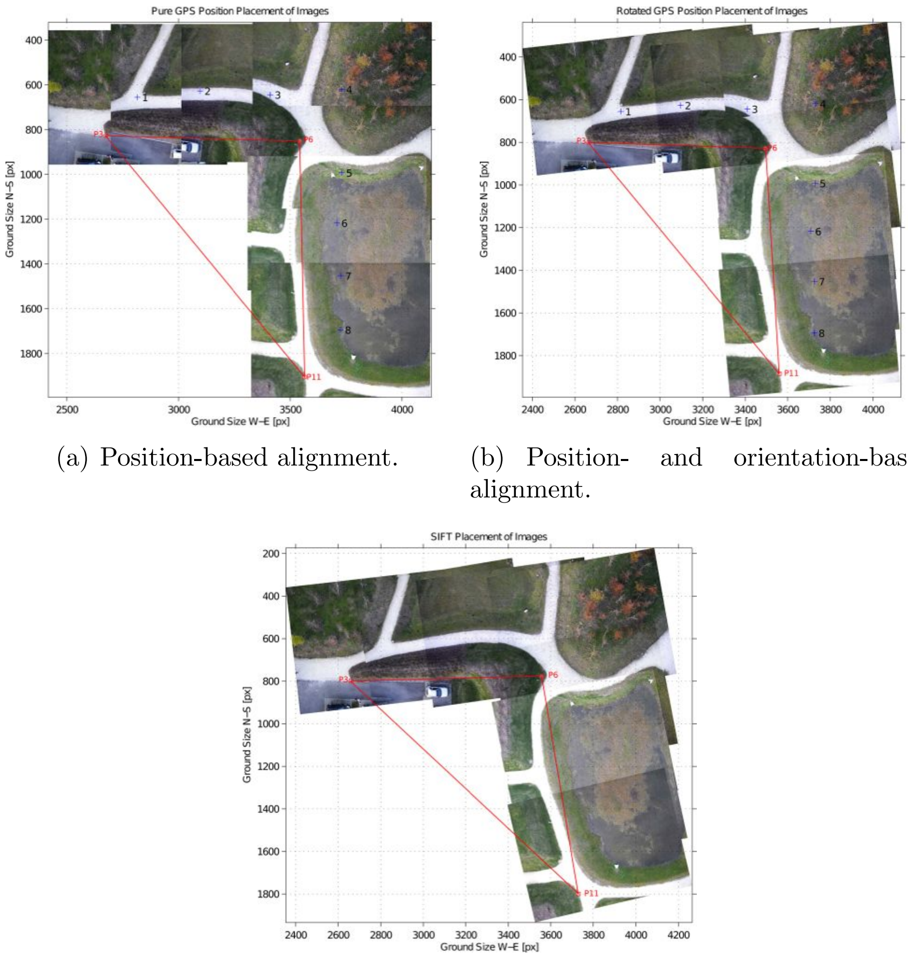
where the scalar s represents the scale and θ represents the rotation angle.

Image-based alignment

Image-based alignment can be categorized into (i) pixel-based, and (ii) feature-based methods (cp. Section 2.3). The idea is to find transformations $\tilde{\mathbf{H}}_{I_i,I'_i}$ and consequently the position of each new image which maximizes the quality function

$$\lambda(\mathit{Merge}(O_{i-1}, I'_i)). \tag{4.5}$$

This quality function can be constructed using pixel-based or feature-based approaches. The pixel-based approaches are computationally more expensive because the quality function is computed from all pixels in the overlapping parts of two images. Feature-based approaches try to reduce the computational effort by first



(a) Position-based alignment.

(b) Position- and orientation-based alignment.

(c) Image-based alignment using SIFT features.

Figure 4.1: Results of basic image mosaicking approaches. The red triangle depicts the distances to compute the spatial accuracy. The units are given in pixels.

extracting distinctive feature points and then match the feature points in overlapping parts. Depending on the chosen degree of freedom the resulting transformation ranges from a similarity transformation to a perspective transformation. The quality function λ will be explained later in this section and will be used for the evaluation purposes in Chapter 6.

The benefit of this approach is that the generated overview image is visually more appealing. On the other hand, the major disadvantages are that the search space grows with the number of images to be mosaicked and the images may get distorted (cp. Figure 3.1).

Hybrid approach

By hybrid approach we propose a combination of metadata-based and image-based methods. The idea is to first place the new images based on the camera's position and orientation information on the already generated overview image. In the next step, we use image-based methods to correct for inaccurate position and orientation information and at the same time improve the visual appearance. Since we already know the approximate position of the image from the camera's position we can reduce the search space significantly. Thus, we split the transformation $\tilde{\mathbf{H}}_{I_i,I'_i}$ into two transformations whereas the $\tilde{\mathbf{H}}_{pos,I_i,I'_i}$ represents the transformation based on the camera's position and orientation and $\tilde{\mathbf{H}}_{img,I_i,I'_i}$ represents the transformation which optimizes the alignment using the image-based method. We find the transformations $\tilde{\mathbf{H}}_{pos}$ and $\tilde{\mathbf{H}}_{img}$ which maximize the quality function

$$\lambda(Merge(O_{i-1}, I'_i)),$$
$$\text{where} \quad \begin{cases} I'_i = T(I_i, \tilde{\mathbf{H}}_{pos,I_i,I'_i}, \tilde{\mathbf{H}}_{img,I_i,I'_i}) \\ pos \in \{(u,v)|u \in [x_{min}, x_{max}], v \in [y_{min}, y_{max}]\} \\ \lambda(O_n) = \alpha \cdot \lambda_{spat}(O_n) + (1-\alpha) \cdot \lambda_{corr}(O_n) \end{cases}, \tag{4.6}$$

and T is a function which transforms image I_i to image I'_i based on combination of $\tilde{\mathbf{H}}_{pos}$ with $\tilde{\mathbf{H}}_{img}$. The combining process is done by using simulated annealing optimization method [30] and search for the target transformation which maximizes our quality function λ. Although simulated annealing usually performs slow, we exploit the image pyramids to compensate for this flaw. The total quality function λ is a weighted combination of λ_{spat} and λ_{corr} ($0 \leq \alpha \leq 1$). λ_{spat} represents the accuracy of spatial distances while λ_{corr} shows the correlation in areas of overlapping images, which is a measure for the seamlessness of the mosaicking. The weight α is set based on application. In other words, a large value for α emphasizes on preserving the relative distances, while a small value emphasizes on visual appealing of a mosaic.

We limit the search space to a reduced set of possible positions based on the expected inaccuracy of position and orientation information. The total error range in the hybrid approach defines the search space in order to find the estimated position. By estimating the appropriate image position we compensate for the total error (GPS and camera tilting errors). Figure 4.2 helps to understand this concept better. We search inside this possible error range to find the best estimated position which maximizes our quality function best. In fact considering a case without any GPS error and considering all views completely nadir, the hybrid algorithm will be reduced to a simple position-based approach.

With this proposed approach we can generate an appealing overview image without significant perspective distortions and at the same time maintain the relative distances in the georeferenced overview image. Moreover, this approach can cope with inaccurate position and orientation information of the camera and thus avoid stitching disparities in the overview image.

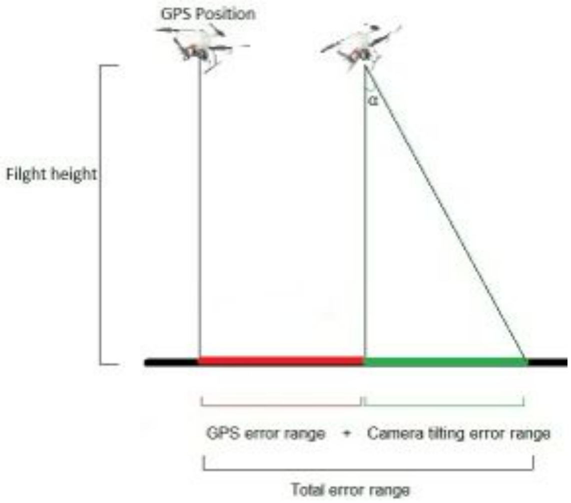


Figure 4.2: The red line shows the GPS error range (the real position is in this range). The green line shows the tilting error range. The sum of this two errors give us the total positioning error

4.2 Loop-independent mosaicking

In this section we survey thoroughly the problem of orthorectified and incremental image mosaicking of a sequence of aerial images in absence of loop(s) (cp. Section 2.3.3) in the image sequence. Since the metadata (acquired by IMU or GPS sensors) in small-scale UAVs are not reliable, they are only used to mitigate the mosaicking errors. Most of other approaches have been exploiting the global optimization to distribute the accumulating error (cp. Section 2.3.4). Without loop in the image sequences, no global optimization can be performed. However, the resulting mosaic can be improved if the errors are diminished by studying their sources. Mostly the UAV aerial image mosaicking is affected by the following three important sources of error:

1. a weak homography as a result of using unleveled ground control points (GCPs) for image registration,
2. a poor camera calibration and image rectification, and
3. deficiency of a well-defined projection model (cylindrical, planar, etc.) and consequently an inappropriate transformation model.

We investigate the influences of using a depth map to find the features from the same plane, geometric distortion correction and combining the appropriate choice of projection and transformation model for the mosaicking. We further quantify the improvement of orthorectification in mosaics by mitigating those errors and demonstrate the improvement on real-world mosaics.

4.2.1 Problem definition

Imagine a case where we want to generate an incremental mosaic of consecutive images taken by UAVs without any loop, knowing the typical pairwise mosaicking explained in Section 2.3. The challenge is how to preserve the orthorectification as much as possible without exploiting any metadata (e.g., GPS or IMU). Consider O_n as the overview image of the target area given a set of n consecutive images $\{I_i | i = 1 \dots n\}$. The overview image can be iteratively constructed the same way as explained in Equation 4.1.

This mosaicking can be described as an optimization problem, in which we need to set the parameters in a way that it maximizes our orthorectification quality function η. One way of constructing such a quality function is using a metric which evaluates the deformation of an image in different directions (horizontal, vertical and diagonal) compared to a reference image,

$$\eta = \frac{4}{\sum_{i=1}^{4} \frac{\max(l_i, l'_i)}{\min(l_i, l'_i)}}, \tag{4.7}$$

where l_i are the length of width, height, and two diagonals of each target image, and l'_i are the length of width, height, and two diagonals of the reference image. In fact, by this metric we calculate the harmonic mean of horizontal, vertical, and two diagonal deviation ratios.

In our work we combine different existing pairwise stitching methods and compare the resulting mosaics in terms of relative distances. Hence, we decide how to set the parameters to obtain the optimal result. Note that although we narrowed down the scope of our scenario, it is possible to simply merge the result with other approaches such as using metadata or global optimization. Though the global optimization methods are more efficient when either there are more than two viewpoints for most of the regions or in existence of a loop in the image sequences.

4.2.2 Major sources of error in pairwise mosaicking

In order to use our metric and compare different mosaics we need a known and well-defined ground truth. For illustration we lined up a set of printed chess-board patterns plus some non-planar objects that we put over and around those patterns along the scene. Then we use a camera with fixed custom settings (e.g., in our case focal length= $28\,mm$, exposure time= $1/500\,s$) and take consecutive images manually from top view with approximately 70% of overlap. By setting a low focal length and consequently a wider angle of view we increase the overlap ratio which leads to more matched feature and inliers. But note that in this case we also encounter a higher geometric distortion. In this way we can simulate the imaging from UAVs to a good extent.

We tested different existing algorithms and parameters that are used for image mosaicking such as internal geometric distortion correction algorithms, feature extraction methods (SIFT, SURF, and Harris corner) with different parameters, projection and transformations models, and manual feature selection. Among all, there are three main parameters that affect the pairwise aerial image mosaicking more than the others namely

1. using unleveled features or GCPs for image registration,

2. internal geometric distortion, and

3. choice of projection and transformation model.

In the following we discuss these parameters and quantify them based on our simulation data-set. In an ideal mosaic all chess-boards should have the same size and shape ($\eta = 1$).

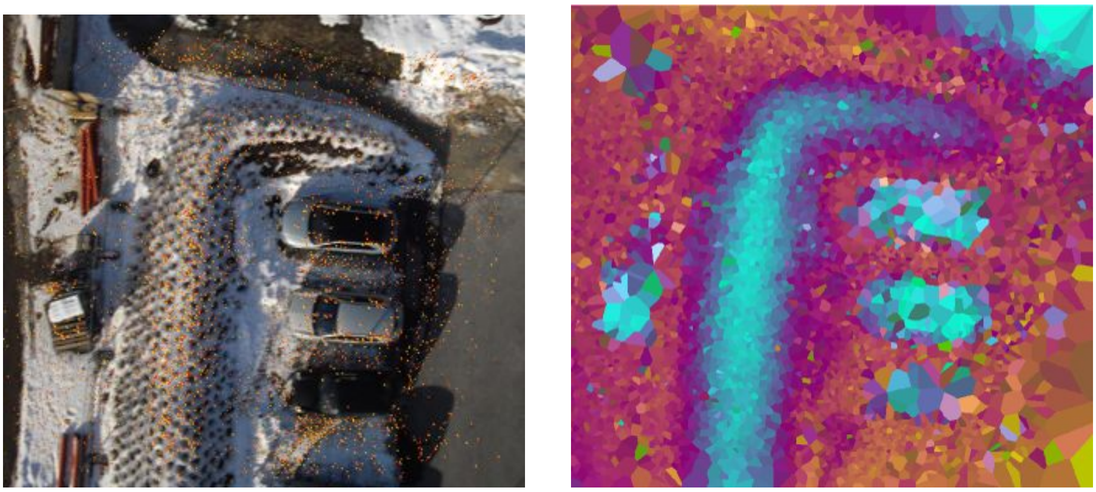
Using unleveled features for image registration

Most of the existing mosaicking algorithms are originally built for panorama construction which consider all images are taken almost from the same spot. In this case the depth variation of the scene is not a problem (except small motions of camera or failure to rotate the camera around its optical center, which is usually handled by a parallax removal algorithm [64]). But in our aerial imagery scenario we take the images from significantly different points of view. As a result, non-planar features produce a disparity when matching features from corresponding images. The disparity vector $\mathbf{d}$ of each transformed feature implies the vector from the expected feature point toward the estimated feature point $\mathbf{d} = (x - \hat{x}, y - \hat{y})$.

These disparities impact the transformation estimation procedure as explained in Equation 2.2. To reduce this effect we need to extract the depth information to extract only the features from the same elevation level which later will be used for image homography. Some depth map construction algorithms use the whole image information (pixels), but we just use the displacement of feature points to speed up the process. Sample disparity vectors from a set of stereo images taken by a UAV are shown in Figure 4.3(a). In order to visualize the corresponding information from these disparity vectors we depict a rough depth map in Figure 4.3(b). The false-color depth map image, $\mathbf{DM}$, is constructed using

$$\begin{aligned}
&\text{Red component} \;\;\; = \mathbf{DM}(x_{n_i}, y_{n_i}, 1) = \tfrac{d_{x_i} - \min_i d_{x_i}}{\max_i d_{x_i} - \min_i d_{x_i}},\\
&\text{Green component} = \mathbf{DM}(x_{n_i}, y_{n_i}, 2) = \tfrac{\|\mathbf{d}_i\| - \min_i \|\mathbf{d}_i\|}{\max_i \|\mathbf{d}_i\| - \min_i \|\mathbf{d}_i\|},\\
&\text{Blue component} \;\;\; = \mathbf{DM}(x_{n_i}, y_{n_i}, 3) = \tfrac{d_{y_i} - \min_i d_{y_i}}{\max_i d_{y_i} - \min_i d_{y_i}},\\
&\mathbf{d}_i = (d_{x_i}, d_{y_i}) = (x_{n_i} - \hat{x}_{n_i}, y_{n_i} - \hat{y}_{n_i}),\\
&\overline{\mathbf{x}}_{n_i} = (x_{n_i}, y_{n_i}, 1),\\
&\overline{\mathbf{H}_{I_n,I_m}\tilde{\mathbf{x}}_{m_i}} = (\hat{x}_{n_i}, \hat{y}_{n_i}, 1),
\end{aligned} \tag{4.8}$$

where the last two equations represent the augmented vectors to convert the homogeneous coordinates back to the Cartesian coordinates; $\mathbf{d}$ represents the disparity



(a) Disparity vectors shows the displacement of transformed feature points from their expected positions.

(b) A rough depth map is depicted by interpolating the disparity vectors.

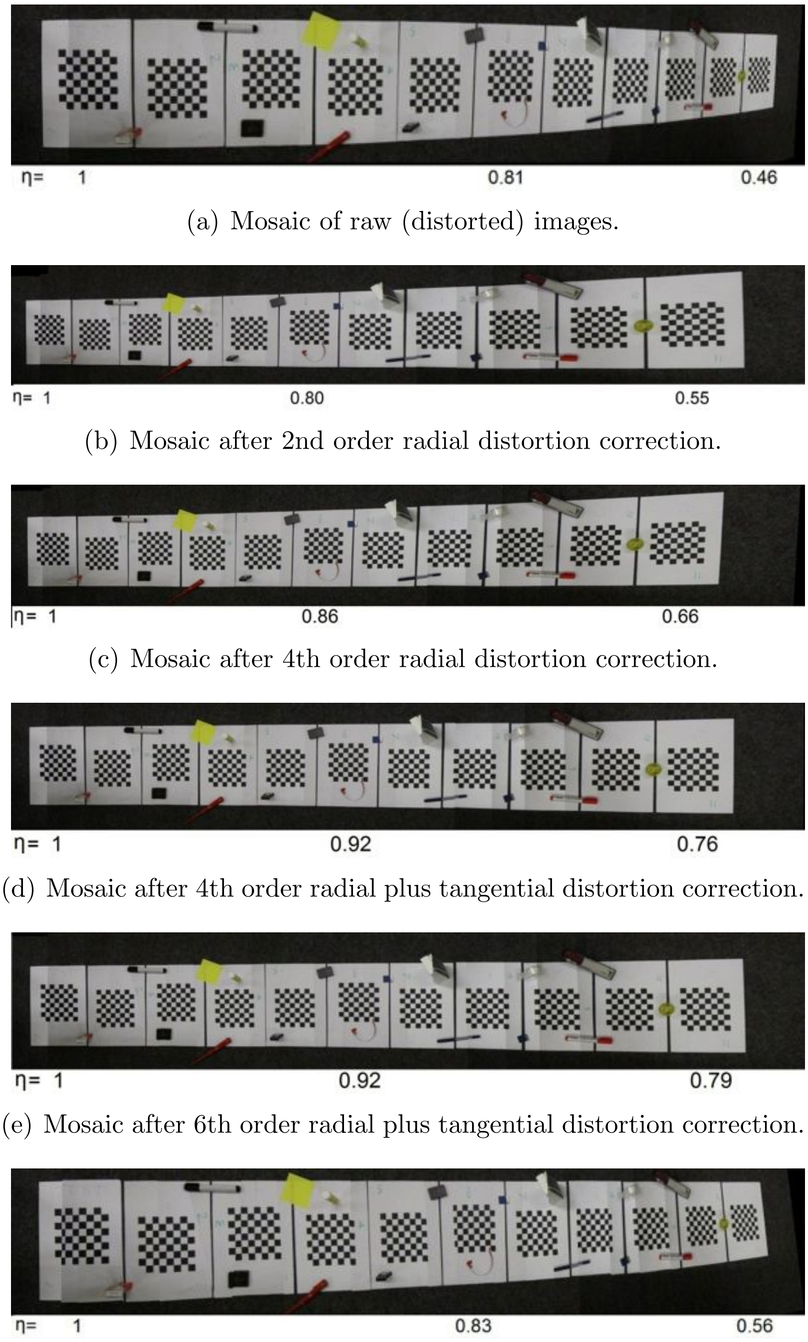
Figure 4.3: Depth information from stereo vision.

vector as displacement between point $\mathbf{x}_n$ and its estimated position $\hat{\mathbf{x}}_n$; i is the index of the inliers among all corresponding feature points. All color components are normalized to fit in the image intensity range. In other words, the normalized x and y components of the disparity vector are used as the red and blue components of the depth map, and the green component of the depth map is the normalized magnitude of the disparity vector. The missing pixels of the depth map image is estimated by interpolation.

We remove features with magnitude of disparity vector $\|\mathbf{d}_i\|$ larger than a threshold ϵ_d. This threshold varies based on height variation of the objects on the ground and flying altitude. In our scenario, we calculate this threshold (in pixels) as follows:

$$\epsilon_d = \left\lceil \frac{\text{maximum height variation}}{\text{minimum flight altitude}} \times 50 \right\rceil , \tag{4.9}$$

where the height variation and the altitude are relative to the dominant ground plane and $\epsilon_d \geq 1$. For example, we set the threshold to 5 pixels if we have a maximum height variation of $4\,m$ and the flight height of $40\,m$. Note that at the first glance it might look similar to setting the RANSAC threshold small, but in that case we might also reject some inliers just because of their small displacement which slows down or even fails the convergence of RANSAC, especially in cases with low amount of overlap. Figure 4.4(f) shows a resulting mosaic of our test model without considering the depth information while in Figure 4.4(e) we see the result with taking the depth into account.



(a) Mosaic of raw (distorted) images.

(b) Mosaic after 2nd order radial distortion correction.

(c) Mosaic after 4th order radial distortion correction.

(d) Mosaic after 4th order radial plus tangential distortion correction.

(e) Mosaic after 6th order radial plus tangential distortion correction.

(f) Mosaic after 6th order radial plus tangential distortion correction but no depth consideration.

Figure 4.4: Resulting mosaic of 21 sequential images with different distortion correction and depth consideration parameters. Note that the η values under the first, middle and last chess-board show the corresponding rectification quality.

Internal geometric distortion

In this section we present the influence of different orders of geometric distortion correction (cp. Section 2.1.2) over the resulting mosaic of 21 consecutive images obtained as described for our test scenario. Figures 4.4(a) to 4.4(e) depicts the results under various distortion correction parameters (the depth information is already considered). The pairwise stitching is performed from left to right. This gives us a visual understanding how much the polynomial orders in distortion correction procedure affects the mosaicking.

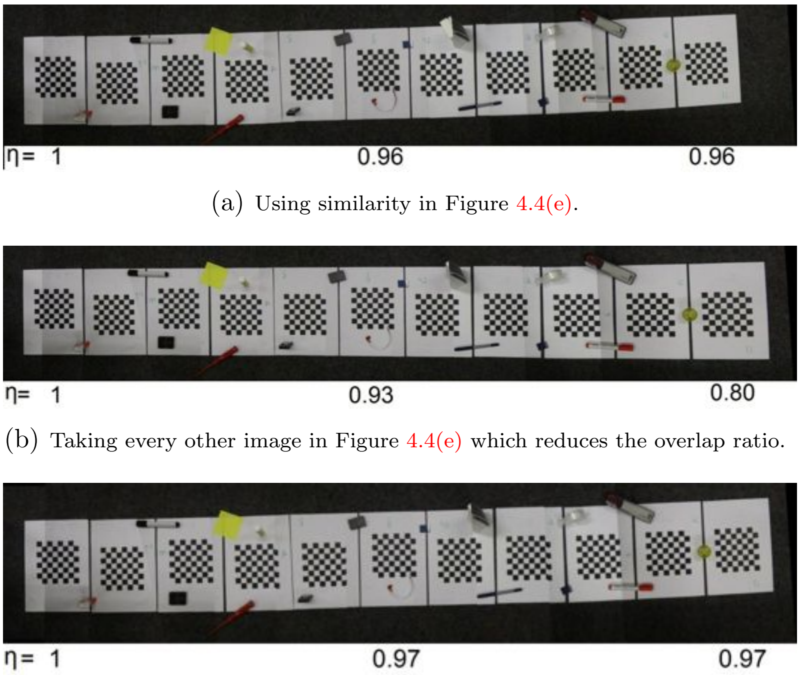
Projection and transformation model

As we mentioned earlier, the planar projection model is an appropriate model for UAV imaging over a plane ground. Choosing the planar model demands a projective transformation to correct the perspective distortion of images taken while the camera was tilted. On the other hand, the projective transformation is quite susceptible to errors and a small deviation will spread after a number of images. Substituting the projective transformation with similarity transformation might help significantly to produce a more orthorectified mosaic, especially in cases in which the first two steps (considering the leveled features by using depth map and correcting the lens distortion accurately) did not manage to restrain the error propagation. The only drawback of using similarity transformation is that it might lead to small seams in pairwise mosaicking which can be ignored if UAV has almost a nadir-view. In Figure 4.5(b) every other image is considered for mosaicking which reduces the overlap ratio. As η values in Figures 4.5(a) and 4.5(c) show, using similarity transformation reduces the deformation.

4.3 Summary on orthorectified mosaicking

In this chapter we first proposed a hybrid approach for mosaicking by combining the metadata with image processing. With this method we improve the quality of a mosaic by georeferencing and maintaining the orthorectification in incremental mosaicking. In each step we initially pose the new image inside the overview image based on the metadata information. Then we adjust and tune the position and orientation of the image within a possible range by using the pairwise image mosaicking explained in Section 2.3. Sample results and the quality metric is presented in Section 6.1.

We also have shown without using the metadata it is possible to improve the quality of mosaicking. We studied the sources of errors in the process of pairwise and loop-independent mosaicking. We discovered that using leveled features for homography estimation, accurate lens distortion correction and choosing the appropriate transformation and projection model produce more accurate and orthorectified mosaics. Mosaics in Figure 4.6 are constructed with this method. Both mosaics include

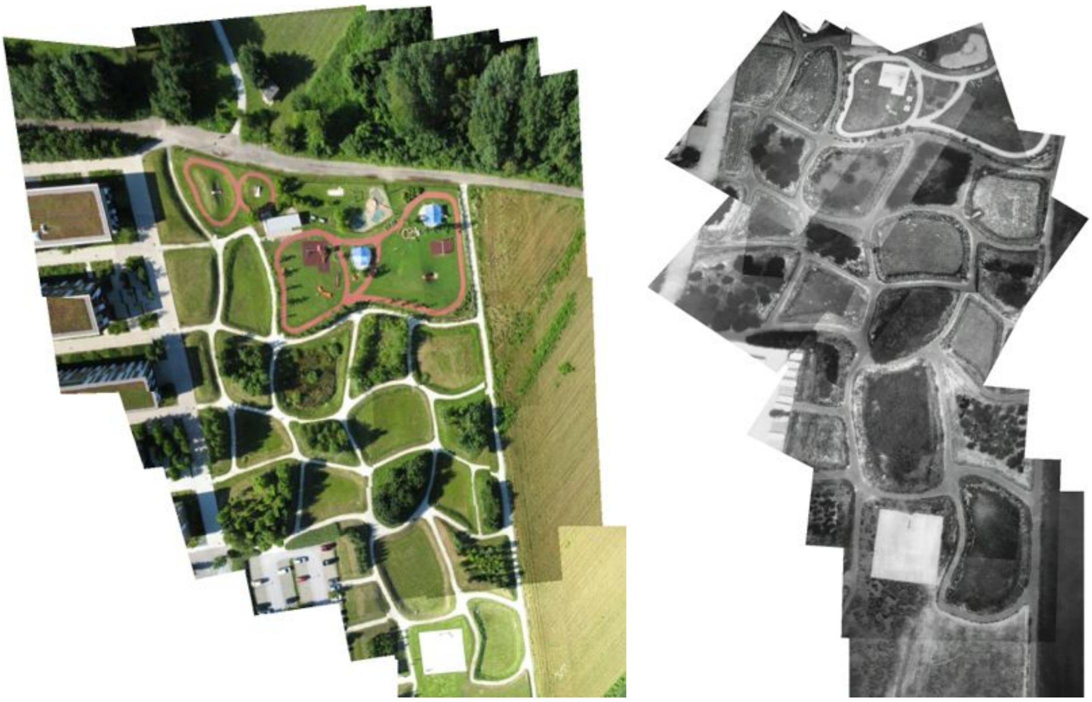


(a) Using similarity in Figure 4.4(e).

(b) Taking every other image in Figure 4.4(e) which reduces the overlap ratio.

(c) Taking every other image in Figure 4.4(e) and using similarity.

Figure 4.5: Resulting mosaic of 21 sequential images with different transformation model. η is the rectification quality.



(a) Mosaic of 24 visual images.

(b) Mosaic of 14 thermal images.

Figure 4.6: Loop-independent, incremental and pairwise mosaicking results.

loops in their image sequences, although no optimization method is used (mosaicking is loop-independent). A seamless mosaic by this method means that the first and the last image of the loop are aligned well, which is a result of orthorectified mosaicking. A quantitative comparison and sample improved results are presented in Section 6.2.

CHAPTER 5 Multispectral mosaicking

Over the years, the use of thermal cameras and capturing images at different spectral bands has been expanded. Initially mainly used by military and government agencies for surveillance and security, the thermal camera technology has now migrated to many other exciting areas. A rapidly developing area for thermal cameras is multispectral aerial imagery based on small UAVs (cp. Figures 1.1(b) and 1.1(c)). A fundamental problem in multispectral imagery is the registration of the individual images captured of the same scene but at different spectral bands. In case of fixed sensors or camera settings, registration is easier because the intrinsic and extrinsic parameters of the cameras can be determined. Hence the relative geometric mapping between the cameras is known, and the transformation for the image alignment can be computed based on purely geometric information. Registration becomes more challenging when this relation is not known. In this case, information of the images must be exploited, i.e., registration basically relies on the identification of robust correspondences between individual images.

In this chapter we first propose a general method for the lens distortion correction of thermal cameras. Then we present three robust methods of feature extraction for the purpose of interspectral image registration. Our first method extends existing robust feature extraction methods to make it wider applicable for interspectral image registration. The other two methods extract additional features in presence of multiple pairs of thermal and visual images. Both methods are advantageous for the registration of low-altitude aerial images.

5.1 Thermal lens distortion correction

Compensating the lens distortion is a very important prerequisite for successful multispectral image processing. While the lens distortion correction is a well established field in visual images (e.g., [4, 6]), it is mostly uncharted territory for thermal imaging. The only thermal lens calibration method we could find was proposed by Rudol and Doherty [51]. Their approach is hard to re-implement since they make use of specially built and not-readily available components. In contrast, the approach proposed by us require only a printed calibration pattern and IR radiation (most conveniently generated by a heat lamp).

Using methods already developed for the lens calibration of visual cameras might not be useful for thermal cameras because:

- Patterns used for visual cameras do not emit any IR radiation per se, and consequently may not be visible to thermal sensors,
- Lenses used for thermal cameras are different than ones used for visual imaging (considering the shape, material, etc.).

In order to take advantage of the readily available calibration tools we adopt the work-flow of the visual lens calibration method which consists of the following steps:

1. acquire a picture (or pictures) of a calibration pattern,
2. detect relevant features of the pattern, and
3. calculate intrinsic and extrinsic parameters of the lens based on features extracted in previous point, mathematical models of the lens and known properties of the used pattern (such as straight or parallel lines, right angles, etc.).

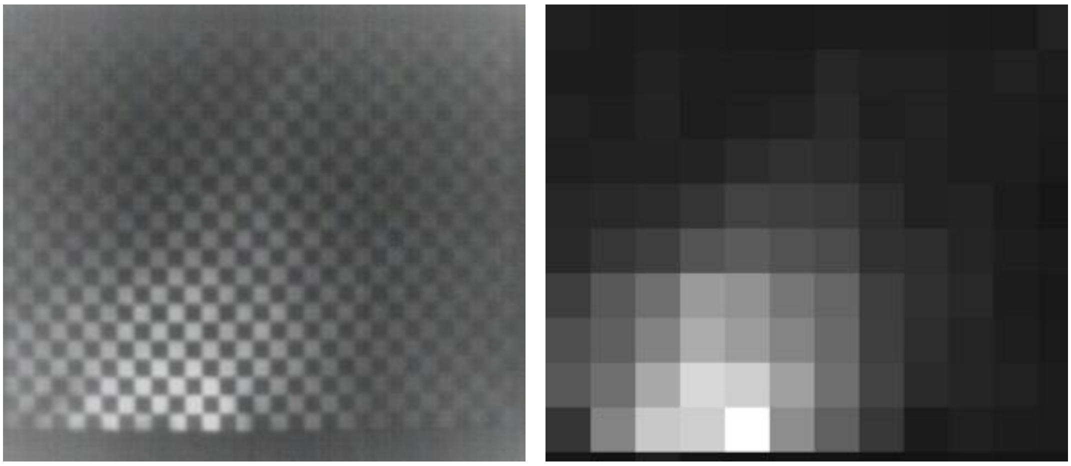
The calibration method presented here replaces the steps 1 and 2 that are tailored to thermal imaging. For the remaining steps we can apply the visual lens calibration methods based on chess-board patterns (e.g., camera calibration toolbox for matlab [4]).

5.1.1 Proposed method setup

In our method we aim to make the conventional chess-board pattern visible for the thermal cameras by radiating the IR light over the pattern. In a chess-board pattern, the black squares absorb more radiation and therefore heat up more than the white squares. Hence, we are able to sense the reflected IR radiation by thermal sensors. In first step we heat up a normal chess-board pattern printed on an A0 [1] paper with IR radiation. We use a IR radiating lamp which can be precisely controlled. The main problem with the IR lamps is that their field of operation is not wide enough to heat up the whole pattern uniformly. The reason that we have chosen such a big pattern is that the minimum focus distance of our thermal lens is $2\,m$ (For detailed description of the test rig see Section 5.1.4).

The following idea has been implemented to resolve that issue. A fixed thermal camera takes series of images while the IR emitting lamp is moved across the pattern. Then images are analyzed and the final image is assembled from pieces of those input images with highest contrast. This procedure will be explained in detail in the Section 5.1.2.

[1] dimensions: $841 \times 1189\,mm$ (ISO 216)



(a) A sample frame out of all set of frames.

(b) The weight function over the corresponding frame.

Figure 5.1: Extracting the target area.

5.1.2 Maximum contrast image assembly

The input of the assembly is series of images $\{\mathbf{I}_i | i = 1 \ldots n\}$ taken exactly from the same area as explained in the Section 5.1.1. Figure 5.1(a) shows a sample frame of thermal image while the pattern is partially heated up by IR radiation.

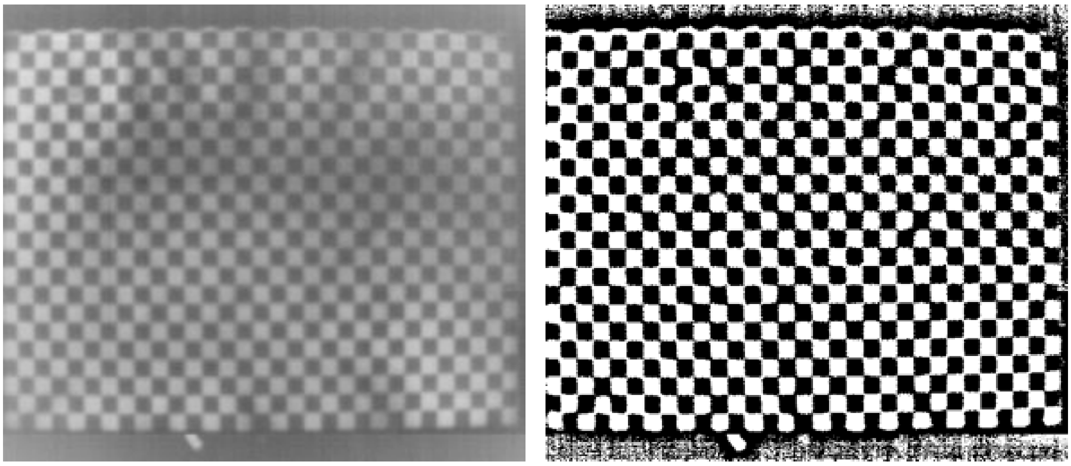
In order to extract the position of the squares in chess-board we perform the following steps over the set of frames:

- First we need to crop a region of each frame which has the required information. In other words, since the IR light heats up only a part of the pattern we need to extract that specific area. As you can see in Figure 5.1(a) inside this target area the chess-board pattern has more contrast and therefore it produces a more clear edge at each square as compared to other regions. Since the bimodality of the chess-board is a good characteristic of this area we use the metric [78] which favors an image intensity histogram with two peaks where intra-peak variance is small and inter-peak distance is large

$$\varphi(C) = \frac{\sigma_{C'} + \sigma_{C''}}{(\mu_{C'} - \mu_{C''})^2}$$

$$\text{where} \quad \left\{ \begin{array}{l} C' = \{C(x) \mid C(x) < \mu_C\} \\ C'' = \{C(x) \mid C(x) > \mu_C\} \end{array} \right. , \tag{5.1}$$

$C \subset I$ is the subregion of I which we measure the bimodality over. $\mu_{()}$ and $\sigma_{()}$ show the mean and the variance, respectively. We partition each image into a grid. The size of the grid depends on the size of the black and white squares. To avoid aliasing we obey the Nyquist sampling rate so that the



(a) Sample integration of 92 individual frames. (b) The result after adaptive thresholding.

Figure 5.2: Constructed integration of frames.

minimum grid size is twice the size of each black or white square. Afterwards we compute the metric over each subregion C_{ij} of the grid. Figure 5.1(b) shows the corresponding weight function which is constructed using

$$\varphi(I) = \begin{bmatrix} \varphi(C_{11}) \cdots \varphi(C_{1n}) \\ \vdots \quad \ddots \quad \vdots \\ \varphi(C_{m1}) \cdots \varphi(C_{mn}) \end{bmatrix}. \tag{5.2}$$

- In the second step we integrate all frames together and construct a single frame with maximum contrast, by

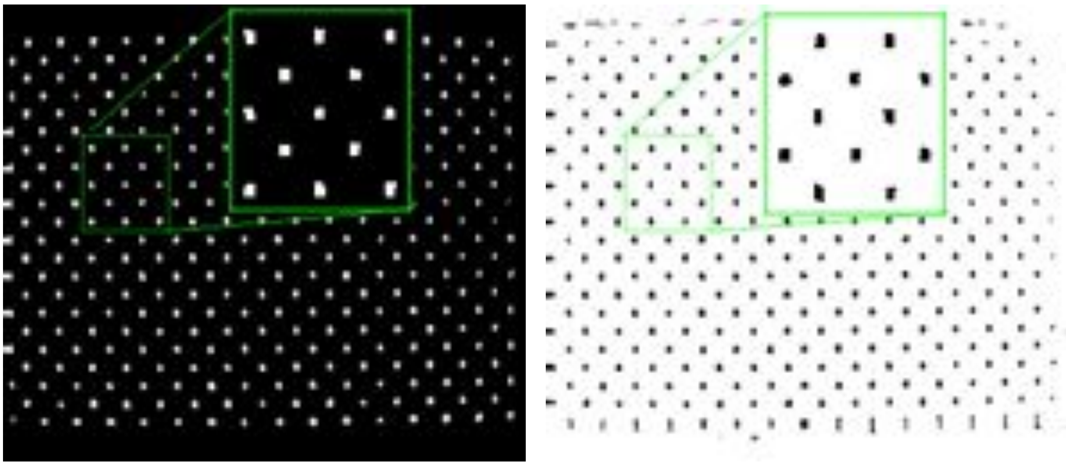
$$O = \frac{\sum_i I_i \cdot \varphi(I_i)}{\sum_i \varphi(I_i)}. \tag{5.3}$$

A sample integration of 92 individual frames is shown in Figure 5.2(a).

5.1.3 Relevant feature detection

Since the images obtained from the assembly described in the previous section are not characterized by very sharp/precise edges, usual visual camera lens calibration methods cannot be yet used. In this stage we describe the process of extracting precise and robust features which can be fed to standard chess-board calibration algorithms.

Using the integrated image from the previous step we continue with the following operations:



(a) The binary image after erosion.

(b) The binary image after dilation.

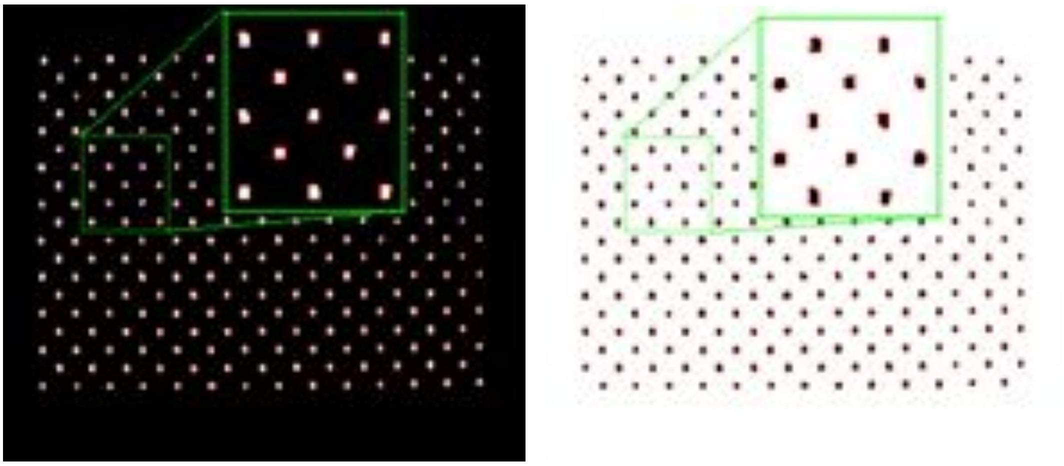
Figure 5.3: Constructed integration of frames.

- Since we know an ideal chess-board pattern should be bimodal, we convert our obtained result to a binary image. Due to the intense variation of the gain level in thermal cameras and also because of the non-uniform IR heating ratio, the constructed integration also has a non-uniform contrast in its different parts. In this case it would be impossible to separate the black and white squares from each other by using a global thresholding. Instead we utilize a quick local adaptive thresholding algorithm proposed by Wellner [72]. The result of thresholding is shown in Figure 5.2(b).

- We define the robust features out of the obtained chess-board pattern. In most of the calibration methods features of choice are corners of the chess-board. As mentioned before the thermal camera may sense the conducted heat from black square to its adjacent pixels and therefore the size of the black squares might become larger than its original size. To avoid this error we mark the center of squares as our features. The center extraction is done by performing first a dilation and erosion over the binary image in a way that we obtain a explicitly distinguished blobs instead of black and white squares as can be seen in Figure 5.3.

- To approximate the center of the squares we need to find the center of the mass of blobs obtained in the previous step. Initially we need to distinguish between individual blobs. In order to do that, we perform a simple contour detection based on method proposed by Suzuki [62]. Desired output of this step is a list of individual contours corresponding to all blobs which is fed to the second step.

- Each contour, which is a list of points defining the outer border of the blob,



(a) The convex hull around the eroded blobs. (b) The convex hull around the dilated blobs.

Figure 5.4: Bounding the blobs by convex hulls.

is processed separately. The smallest convex hulls that include all the points of each contour are calculated. The samples of such convex hulls are shown in Figure 5.4. In these samples, some rows/columns of squares on each margin have been removed by our software to avoid the marginal errors.

- Finally, we calculate the centroid of each blob within its convex hull. Since all pixels have the equal mass, the center of mass is calculated simply by averaging over all pixel locations of each blob: $(\mu(X), \mu(Y))$. The projection of these centers over the chess-board pattern is shown in Figure 5.5.

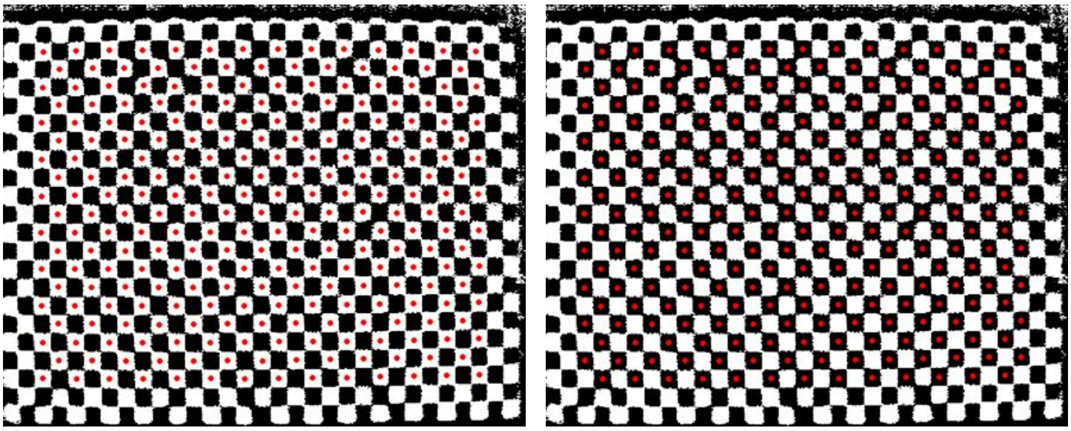
5.1.4 Detailed test rig setup

- Our test were carried out with a thermal camera of the type FLIR Photon 640 [2]. It is a MWIR camera operating in a $7.5 - 13.5\,\mu m$ wavelength range, with a noise equivalent differential temperature (NEdT) of $< 50\,mK$. The resolution of the camera in analog PAL mode is $640 \times 512\,px$. The images have been acquired through analog interface of the camera, by a consumer video grabber card.

- The camera lens is $25\,mm$, $f\backslash 1.4$ with a FOV of 36° x 29°. It has a minimum focus distance of $2\,m$.

- The calibration pattern is a chess-board printed on an A0, $100\,g/m^2$, matte paper, for their superior radiation absorbance/emittance properties. It has

[2] http://www.flir.com/cvs/cores/uncooled/products/photon640/



(a) Centers of eroded squares.

(b) Centers of dilated squares.

Figure 5.5: Square centers visualized as circles.

been attached to a flat wall. The size of the black and white squares are $38 \times 38\,mm$.

- Camera has been mounted on a tripod, $2.5\,m$ from a calibration pattern. It was positioned so the pattern is fitted within the FOV in a way that the optical axis is along the center of the pattern.

5.2 Interspectral registration

In this section we address the robust registration of visual and thermal images captured by different sensors. In this interspectral registration the alignment of the images is typically based on the following steps: (i) extraction of features in the individual images, (ii) matching the corresponding feature points and identifying inliers between those feature points, and (iii) computing the transformations for aligning the individual images. Figure 5.6 shows the schematic description of our work flow for interspectral image registration. Images of different spectrums include rather distinct information. In general, the larger the band difference between the captured images, the more likely the dissimilarity of the features increases. In this section we focus on extracting robust features which can be used for the identification of correspondences. We evaluate the feature point matching of thermal and visual images in general cases and concentrate the registration evaluation on low-altitude aerial images captured by small-scale UAVs. For such small-scale UAVs the number of images and the positions where to capture them are predefined due to limitations in flight time, communication bandwidth, and local processing [43]. In our experience, these images put strong requirements on the registration because of the strong variations in overlap, scale, rotation, point of view, and structure of the

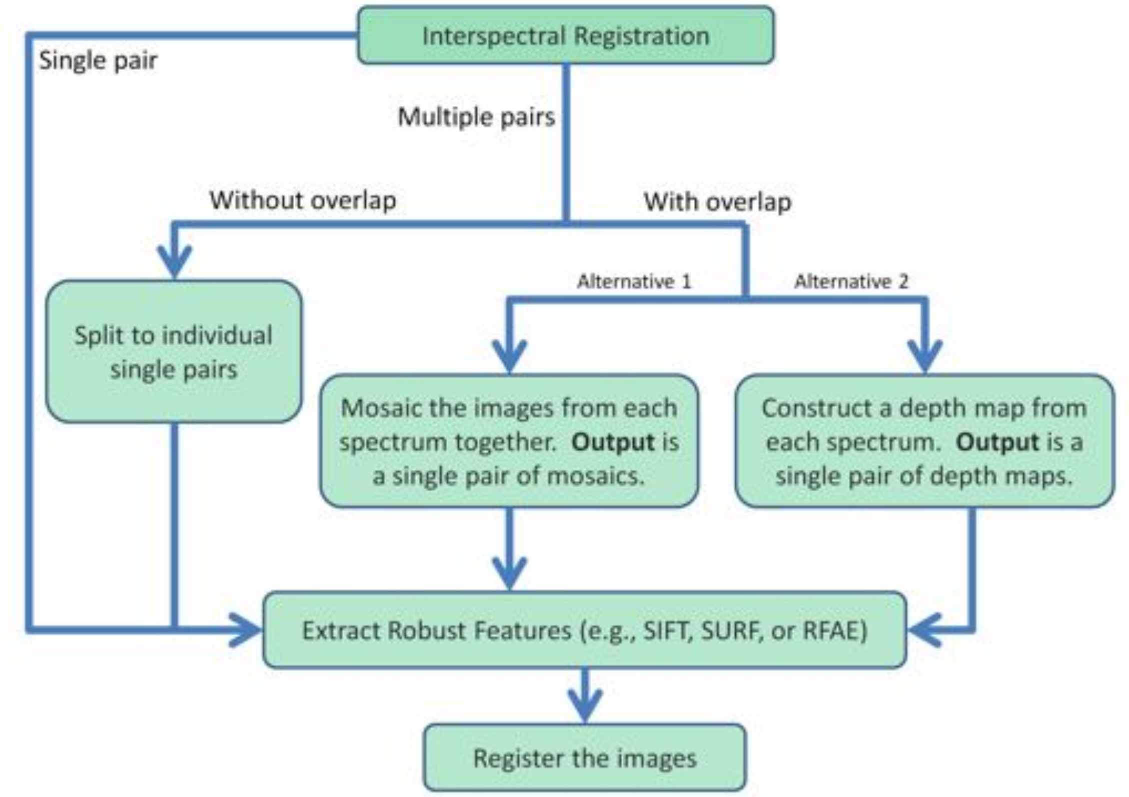
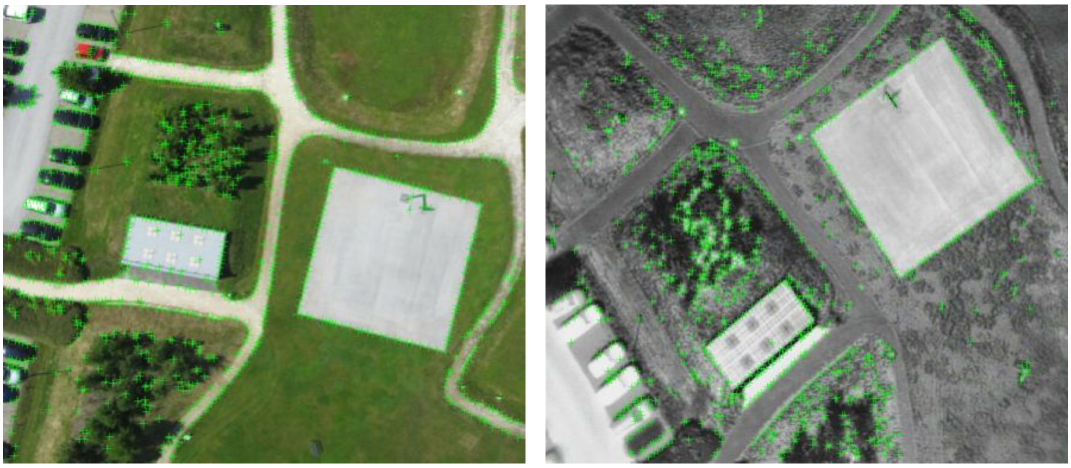

Interspectral Registration
Single pair
Multiple pairs
Without overlap
With overlap
Alternative 1
Alternative 2
Split to individual single pairs
Mosaic the images from each spectrum together. **Output** is a single pair of mosaics.
Construct a depth map from each spectrum. **Output** is a single pair of depth maps.
Extract Robust Features (e.g., SIFT, SURF, or RFAE)
Register the images

Figure 5.6: Interspectral registration pipeline, showing the work flow and different methods we use.

scene.

5.2.1 Analysis of existing feature extraction methods

As previously described, feature extraction is a fundamental step for registration. Although most of the conventional feature extraction methods (such as edge detection or corner detection) can be used to identify the mutual information between visual and thermal images, constructing an appropriate descriptor for finding matching pairs is not so simple. As an example, Figure 5.7 shows the utilization of the Harris operator [20] over a pair of visual and thermal images. This figure demonstrates the differences in the corners extracted in the visual and thermal images. Moreover, the correlation-based matching (or fine-tuning) of the corners fails because of the different intensity pattern and the rotation difference between the two images. In general, the task of matching and removing the outliers and finding the homography becomes challenging in presence of relative rotation and scale between images.



(a) Harris corners in the visual image.

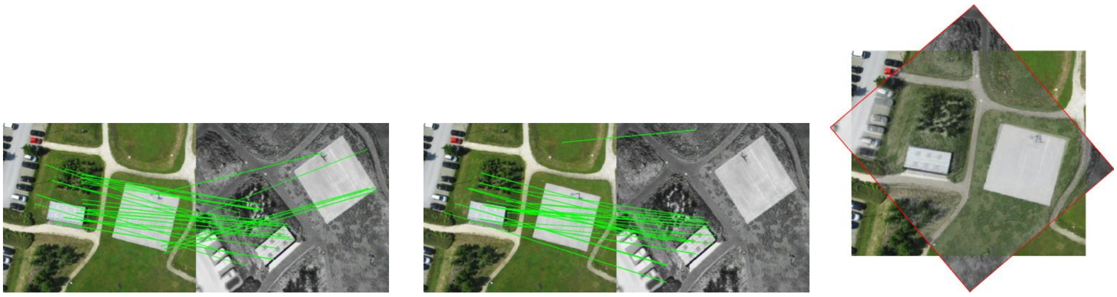
(b) Harris corners in the thermal image.

Figure 5.7: The result of Harris operator over a pair of visual and thermal images.

A multi-scale Harris operator and some other scale-invariant, rotation-invariant, illumination-invariant, and affine-invariant feature extraction and matching methods have been proposed [19, 11, 40, 34] to cope with this limitation. However, the methods with a well-defined robust local descriptor, such as SIFT and SURF, are gaining more attention. Equation 5.4 describes how the SIFT method detects the keypoint locations.

$$\begin{aligned} &D(x, y, \sigma) = (G_(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y), \\ &\text{where} \qquad G(x, y, \sigma) = \tfrac{1}{2\Pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \end{aligned} \tag{5.4}$$

The difference of the Gaussian is used as an approximation for the scale-normalized Laplacian of Gaussian, $\sigma^2\nabla^2 G$. The target keypoints are obtained by calculating the differences between different scales of Gaussian blur over each octave and then by finding the local extrema based on comparing each sample point with its eight neighbors in the current image and its nine neighbors in the scale above and below. Since this method was initially designed for the registration of the images taken from homogeneous sensors, it fails if the parameters are not adjusted in the appropriate way. In other words, when comparing a pair of thermal and visual images taken from the same scene, we may receive matching keypoints at different scale levels even if the images have exactly the same scale ratio. By experiment, we realized that a larger number of scales in the SIFT method improves the registration quality but performs slower because more features needs to be extracted. A detailed study regarding the scales of the SIFT method has been performed by Morel and Yu [41]. If the initial octave is set to -1, feature extraction starts with a double-sized image and consequently obtains more keypoints. In practice, however, the increased image size does not affect the registration quality because there are almost no very



(a) SIFT features with 43 scale levels and Euclidean distance threshold 1.2 for matching.

(b) SURF features with sampling step 1 and initial lobe 1.

(c) Alignment and fusion.

Figure 5.8: Registration of a thermal and visual pair by using SIFT and SURF.

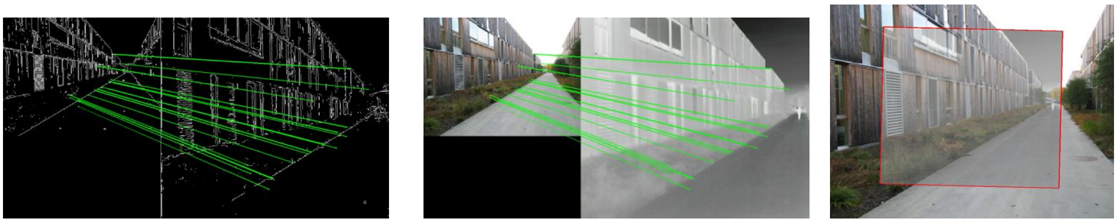
small features in most of the low-resolution thermal images. Nevertheless, setting a lower threshold for the multiplier coefficient of the Euclidean distance of the feature vectors is advantageous for calculating the matching pairs because the matching requirements are not so strict in case of different sensors. Figure 5.8(a) shows an example registration result using the SIFT feature extraction with 43 scale levels and an Euclidean distance threshold of $s = 1.2$ for matching. The default parameter setting does not lead to a successful registration.

SURF analyzes the different scale levels by up-scaling the box filter size rather than iteratively reducing the image size. In this way the performance is highly improved. The keypoint identification is based on an approximation of the determinant of Hessian—instead of the Laplacian of Gaussian in the SIFT descriptor. Figure 5.8(b) shows the same pair of images registered by SURF.

Both methods achieve approximately a 50% successful registration rate by adjusting their parameters based on each scenario. In our data-set we have tested different pairs of satellite images, images of human, images of the nature and surveillance, images taken from UAVs, and images from facade of buildings. We show detailed results of this data-set in Section 6.4.1. Also with the fixed parameters (described in Figure 5.8) both methods achieve a similar performance. Figure 5.8(c) shows a sample aligned and fused result. In all our experiments, we have used RANSAC [12] and least median of squares (LMS) to remove the outliers (among all matched pair-points) and calculate the appropriate similarity transformation between images.

5.2.2 Robust features along the edge (RFAE)

Despite the acceptable performance of SIFT and SURF for interspectral image registration, they have failed in some scenarios in which mutual patterns are clearly available. Apparently both descriptors are inherently designed to emphasize the



(a) Successful registration by extracting the features along the edge.

(b) Overlaying the registration on the original image pair.

(c) Alignment and fusion.

Figure 5.9: Registration of a thermal and visual pair by using scale-invariant features along the edge.

patterns of the gradient changes around a specific keypoint. Note that in different sensors, and more specifically considering thermal and visual sensors, we often record a completely different intensity value for each specific target region. This characteristic affects the matching between the descriptors.

To overcome this problem some authors extract line structures from the images to identify matching points [9, 71, 24]. The main limitation of these methods is that they require a sufficient number of straight lines. Our approach extends this idea and extracts the edge structures in the images. It uses then SIFT or SURF to identify feature descriptors in the extracted binary edge image. This preserves the scale-, rotation-, and illumination-invariant characteristics. In our experiments we have used the Sobel operator as an approximation for the intensity gradient in the images, i.e.,

$$\begin{aligned} \mathbf{S}_x = \tfrac{1}{4}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \qquad & \mathbf{S}_y = \tfrac{1}{4}\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \\ \text{Gradient:} \qquad & \nabla\mathbf{I} \simeq (\mathbf{S}_x * \mathbf{I}, \mathbf{S}_y * \mathbf{I}), \\ \text{Gradient magnitude:} \qquad & \|\nabla\mathbf{I}\| \simeq S(\mathbf{I}) = \sqrt{(\mathbf{S}_x * \mathbf{I})^2 + (\mathbf{S}_y * \mathbf{I})^2}, \\ \text{Gradient direction:} \qquad & \Theta(\nabla\mathbf{I}) \simeq \arctan(\tfrac{\mathbf{S}_x * \mathbf{I}}{\mathbf{S}_y * \mathbf{I}}). \end{aligned} \tag{5.5}$$

To extract the edges, the resulting image is converted to binary by a cutoff threshold. Since the approximation of the gradient becomes bimodal, we reduce the sensitivity of the descriptors to the change of gradient in a neighborhood. Finally, the SIFT and SURF operate the difference of Gaussian or determinant of Hessian over this binary image. In case of SIFT the difference of Gaussian along the edges is defined as

$$D_E(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * B \circ S(I(x, y), \theta), \tag{5.6}$$

where B is the binary operator based on the threshold θ and S is the Sobel operator constructed as explained in Equation 5.5. Figure 5.9 shows the registration of a pair of thermal and visual images taken from a building facade using our method. Note that the registration using merely the SIFT or SURF features failed. A remaining question is how to define the appropriate threshold, θ, for the conversion to the binary edge image. This threshold can be estimated by a statistical analysis of different sensors or different image types. In our experiments, we extract the edges with three different threshold values ($\theta \in \{0.2, 0.4, 0.6\}$). We choose a pair (one threshold for the thermal image and one threshold for the visual image) for registration which maximizes our quality metric (cp. Section 6.4.1).

5.2.3 Interspectral registration by multiple thermal-visual image pairs

So far we have considered the general case of registering thermal and visual images. In this section we focus on registering low-altitude aerial images captured by small-scale UAVs. Due to the payload limitations these aerial robots are typically not able to carry both type of cameras. Figure 1.1 shows two UAVs with different thermal camera models mounted. For these scenarios we improve the robustness of the registration by two methods. The first method exploits entire visual and thermal mosaics. The second method uses depth information to extract additional features for the registration.

Registration of mosaics.

As previously discussed, the robust features along the edge (RFAE) method does not always improve interspectral registration. Figure 5.10 shows an example where RFAE did not improve the registration of two image pairs (i.e., I_{V_1} with I_{T_1} and I_{V_2} with I_{T_2}). The reason is that there are insufficient salient border lines and edges which are visible in both image types. As indicated with green lines in Figure 5.10, one pair (I_{V_1} with I_{T_1}) can be weakly registered with 7 SURF feature matches.

Here we present a new method to exploit the image mosaics to strengthen the interspectral registration. The mosaicking of aerial images taken from an identical sensor is based on the homography $\mathbf{H}$ corresponding to the perspective transformation between each pair of images (cp. Section 2.3). Thus, pairwise registration can be seen as an initial step for mosaicking. Registration within a specific spectrum (identical sensor) is typically more robust and can be achieved even with a limited pairwise overlap. As shown in Figure 5.10, the visual image I_{V_2} is transformed to the coordinates of the visual image I_{V_1} by homography $\tilde{\mathbf{H}}_{I_{V_2},I_{V_1}}$. Similarly, the thermal image I_{T_2} is transformed to the coordinates of the thermal image I_{T_1} by homography $\tilde{\mathbf{H}}_{I_{T_2},I_{T_1}}$. By knowing one of the interspectral registration parameters, for example the corresponding pair points between images I_{V_1}, I_{T_1} shown as $R(\tilde{\mathbf{x}}_{V_1}, \tilde{\mathbf{x}}_{T_1})$, we can calculate the corresponding pair points between images I_{V_2}, I_{T_2} by

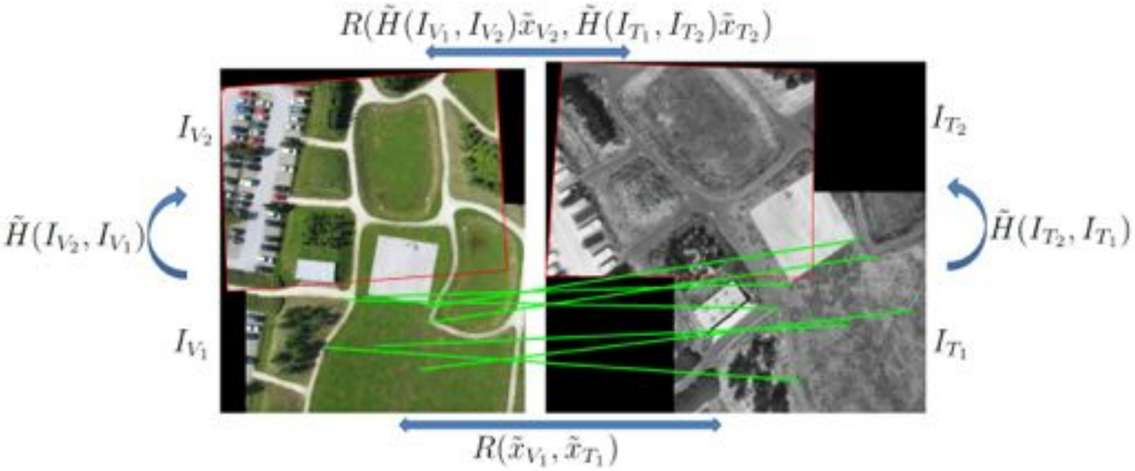
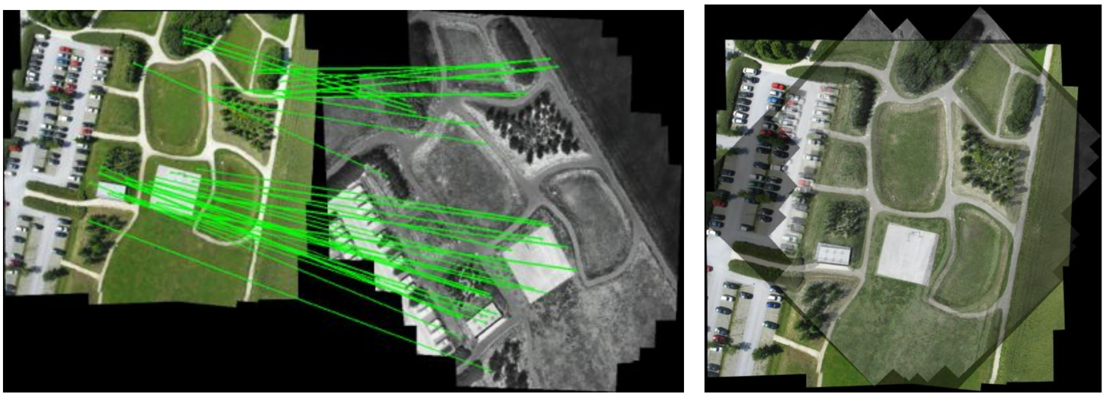

$R(\tilde{H}(I_{V_1}, I_{V_2})\tilde{x}_{V_2}, \tilde{H}(I_{T_1}, I_{T_2})\tilde{x}_{T_2})$
I_{V_2}
$\tilde{H}(I_{V_2}, I_{V_1})$
I_{V_1}
I_{T_2}
$\tilde{H}(I_{T_2}, I_{T_1})$
I_{T_1}
$R(\tilde{x}_{V_1}, \tilde{x}_{T_1})$

Figure 5.10: Interspectral registration by using multiple pairs.

$$R(\tilde{\mathbf{x}}_{V_2}, \tilde{\mathbf{x}}_{T_2}) = R(\tilde{\mathbf{H}}_{I_{V_1}, I_{V_2}} \tilde{\mathbf{x}}_{V_2}, \tilde{\mathbf{H}}_{I_{T_1}, I_{T_2}} \tilde{\mathbf{x}}_{T_2}), \tag{5.7}$$

where $\tilde{\mathbf{x}}, \tilde{\mathbf{H}}$ are the points and the homography in homogeneous coordinates, respectively.

The interspectral registration between large mosaics can be done in two different ways. The first way is to generalize the approach shown in Figure 5.10 over multiple pairs. No matter if the registration fails in some pairs, the mosaics can be registered as long as some of the thermal and visual images are registered. However, this method needs to consider all corresponding points, both within the homogeneous and the heterogeneous image types, for the global optimization. In other words, we need to find out the homographies which minimize the least mean squares (LMS) of the disparity error between all pair points. This increases the complexity of the homography estimation and the mosaic construction. Points that are considered in multiple image pairs (when more than two images overlap) will be over-weighted in this optimization. The knowledge of corresponding images, i.e., images with are supposed to have some overlap, is also required for this method. The second way is to first mosaic all the images from the same sensor and then register the two final mosaics together. The thermal and visual mosaics shown in Figure 5.11(a) are registered with this approach. The drawback here is that handling large image mosaics and large number of corresponding points is computationally expensive. In addition, errors in mosaicking homogeneous images affect the interspectral registration accuracy.



(a) Registered correspondences.

(b) Overlayed and fused images.

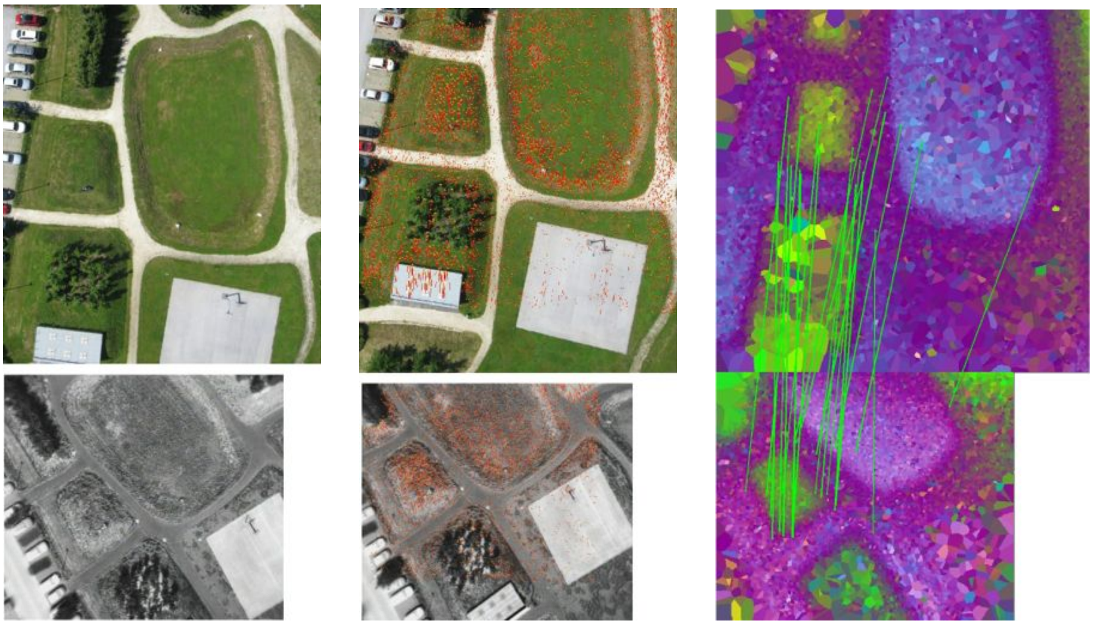
Figure 5.11: Interspectral registration of two mosaics each constructed from 25 individual images.

Exploiting the 3D structure

In scenarios where UAVs provide sparse pictures from different points of view, we can exploit methods from stereo vision to extract depth information. In this section we describe how depth information helps for the registration of thermal and visual image mosaics. As explained in Section 4.2.2, we reveal depth information of a scene by using stereo images of the same scene taken from different points. We first calculate the disparity vectors from the displacement of all feature pairs in the two stereo images. Figure 5.12(b) depicts these disparity vectors as the displacement between two feature points after the alignment. Since this requires stereo images, we can only obtain the disparity vectors over the overlapping area of images taken from different positions. The magnitude of a disparity vector corresponds to the relative height difference of the corresponding feature point. The direction of the vector determines whether the feature point is below or above the average altitude. This helps us to construct a rough depth map as shown in Figure 5.12(c). The false-color depth map image, **DM**, is constructed as explained in Section 4.2.2.

By extracting the depth map of the overlapping area in both thermal and visual image pairs, we are able to register those images by registering their depth map. Regardless of existence of any mutual pattern or similarity between visual and thermal images, the depth information of a target scene provides a consistent mutual information between two image types. An automatic registration based on SURF features is shown by the green lines in Figure 5.12(c). We can generalize the depth map construction from a pairwise depth map to a mosaic depth map. The disparity vectors are constructed as explained in Equation 4.8.

An alternative to this 2D mosaicking is a 3D optimization by a full bundle adjustment and estimating and reconstructing the 3D point positions [59, 16, 66]. The 3D models shown in Figure 5.13 are generated by such 3D reconstruction from

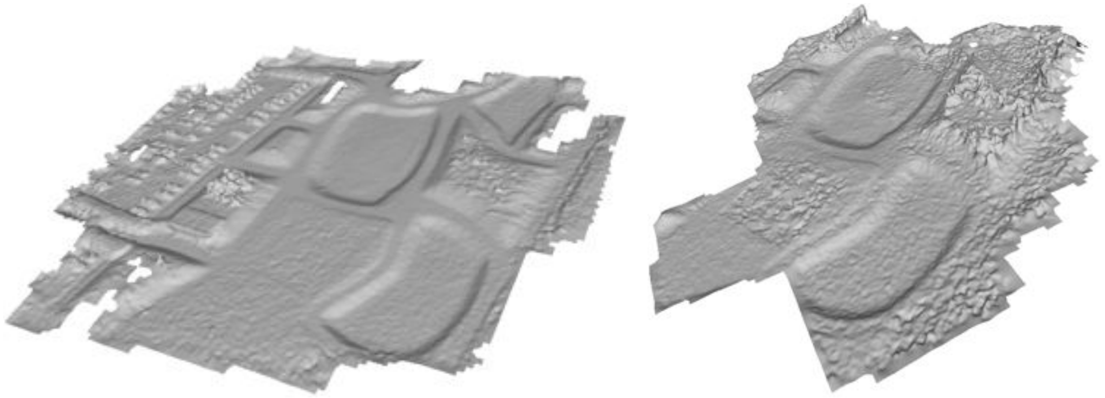


(a) Visual and thermal images taken at initial UAV position.

(b) Disparity vectors depicted in the overlapping area with images taken at a different UAV position.

(c) Computed depth maps of the overlap. The green lines indicate the registration based on SURF.

Figure 5.12: Construction of the depth map from two image pairs by calculating the disparity of the feature points. The upper row corresponds to the visual and the lower row to the thermal images.



(a) 3D model from 25 visual images.

(b) 3D model from 25 thermal images.

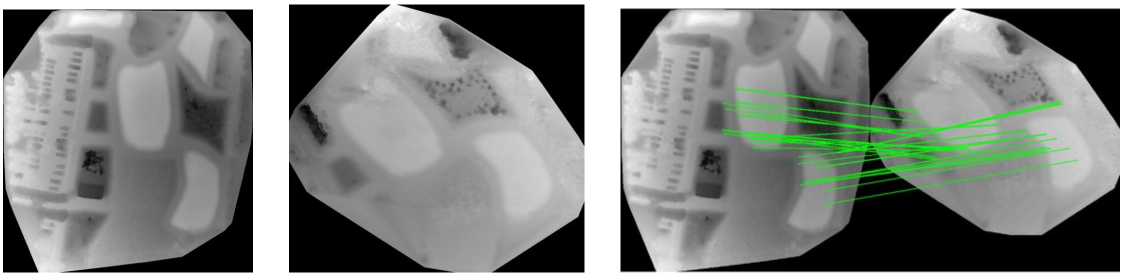
Figure 5.13: 3D model reconstruction of the target area by using a full 3D bundle adjustment.

the same 25 thermal and visual images as used in Figure 5.11. In general, a full 3D model reconstruction achieves higher accuracy with more images. Despite this fact, we often face more challenging scenarios such as sparse images with limited overlap. Furthermore, most of the thermal cameras mounted on small-scale UAVs have a lower resolution as compared to visual cameras and provide an analog image. These cameras often do not have a global shutter and correcting the lens distortion is not straightforward (cp. Section 5.1). Hence, we typically see more noise in the 3D models constructed from thermal images than from visual images. It is therefore difficult to register these 3D models which also may be different in scale and orientation. Most of the existing point cloud or 3D-mesh registration methods considers a high accuracy of the models [13, 56, 60].

Since we are more interested in 2D images and 2D registration, we perform a mapping transformation to convert our 3D models to an equivalent depth map. We first map all 3D points to a 2D plane Π, which is almost parallel to the ground and the camera planes. This is performed by finding $P_{i\Pi}$, the base point of the perpendicular of each 3D point $\mathbf{P}_i = (x_i, y_i, z_i)$, to the plane $\Pi : ax + by + cz + d = 0$ with a normal vector $\mathbf{n} = (a, b, c)$. Second, the distance $d(\mathbf{P}_i, \Pi)$ of the point $\mathbf{P}_i$ to the plane Π is translated to the intensity value of the depth map image,

$$\mathbf{P}_{i\Pi} = \mathbf{P}_i - \frac{ax_i + by_i + cz_i + d}{a^2 + b^2 + c^2}\mathbf{n}, \quad d(\mathbf{P}_i, \Pi) = \frac{ax_i + by_i + cz_i + d}{\sqrt{a^2 + b^2 + c^2}}. \tag{5.8}$$

Figures 5.14(a) and 5.14(b) show such depth maps constructed from the visual and thermal 3D models in Figure 5.13. The 3D reconstruction is much slower compared to the fast depth map construction based on disparity. However, the depth map images as results of 3D reconstruction methods are smoother. Figure 5.14(c)



(a) Corr. depth map of the Figure 5.13(a).

(b) Corr. depth map of the Figure 5.13(b).

(c) Automatic registration done using the SURF features.

Figure 5.14: Extraction of the depth maps from the 3D models of Figure 5.13.

depicts the automatic registration of the resulting depth maps by using the SURF features. One obvious advantage of this registration method for thermal and visual images is the robustness against the image differences and spatial changes, since we register the depth information and not the image details. This is especially useful in cases with a high time difference between two remote sensing activities, such as registering images captured in different seasons of the year.

CHAPTER 6 Results and discussion

In this chapter we extend the discussion and evaluate different methods explained in previous chapters. In Section 6.1 we construct a quality function to evaluate our hybrid approach for incremental mosaicking. The results from this section are mutual works with my colleague Daniel Wischounig-Strucl and are partially published in [79, 49]. In Section 6.2 we depict some improvements as a result of mitigating errors in loop-independent mosaicking. The result of this section is published in [77]. In Section 6.3 we evaluate our method for lens distortion correction of a thermal camera. We also show that in a real scenario by correcting the lens distortion we improve the quality of the interspectral registration. The result of this section is published in [76]. In Section 6.4 we extendedly evaluate the RFAE and other methods for the purpose of interspectral registration.

6.1 Incremental mosaicking

In this section we compare the results of the hybrid mosaicking with the other three approaches in Section 4.1.2. This evaluation mainly focuses on the geospatial accuracy and image correlation which are specified in our quality metric in Section 6.1.1. We further compare the required computation times of all approaches which have been implemented in Matlab on a standard PC running at $2.66\,GHz$.

For the evaluation we use a rectangular round trip mission for which 40 picture-points have been planned (cp. Figures 6.1, 6.2, and 6.3). Images have been captured from a single UAV flying at an altitude of approximately $30\,m$. The overlap among adjacent images is about 60%. However, three images were lost in the real UAV mission (cp. positions B, C, and D in Figure 6.3) which reduces the overlap in these specific areas to approximately 20%. A subset of 8 images (cp. Figure 4.1) is used to compare the mosaicking results of the three approaches explained in Section 4.1.2.

6.1.1 Quality evaluation

To evaluate the quality of the different mosaicking approaches presented in Section 4.1.2, the λ_{spat} and λ_{corr} in Equation 4.6 are defined as follows:

$$\begin{aligned}\lambda_{spat} &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{1+|\frac{d_i-\hat{d}_i}{d_i}|},\\ \lambda_{corr} &= \frac{1}{n}\sum_{i=1}^{n}\frac{1+CC(\mathrm{Mask}(O_{i-1},I'_i),\mathrm{Mask}(I'_i,O_{i-1}))}{2},\\ CC(X,Y) &= \frac{\mathrm{Covariance}(X,Y)}{\sigma_X\sigma_Y},\end{aligned} \tag{6.1}$$

where $\mathrm{Mask}(I_n, I_m)$ represents a part of the image I_n that has overlap with the image I_m, d_i is the actual distance measured between two ground control points, $\hat{d}_i$ is the estimated distance extracted from overview image and m is the number of considered distances. As it can be deduced from the Equations 4.6 and 6.1, λ, λ_{spat} and λ_{corr} are all in the range of $(0, 1]$. In our evaluations we set the weight α used in Equation 4.6 to 0.5. As mentioned in Section 4.1.2, based on application, a large value for α emphasizes on preserving the relative distances, while a small value emphasizes on visual appealing of a mosaic.

6.1.2 Metadata- and image-based approaches

In our evaluation the quality of mosaicking for the first 8 images of the round trip mission $\lambda(O_8)$ is calculated. In order to evaluate the spatial quality $\lambda_{spat}(O_8)$ we chose a triangle, spanning significant points (P_3, P_6, P_{11}) for simplified spatial evaluation in the reduced set of eight images. In Table 6.1 the measured distances $(|\overline{P_3P_6}|, |\overline{P_6P_{11}}|, |\overline{P_3P_{11}}|)$, the resulting spatial quality, and the correlation quality are presented and combined according to Equation 6.1 to a final quality characteristic to compare the presented approaches. As this table shows, the values of $\lambda_{corr}(O_8)$ and $\lambda_{spat}(O_8)$ for the mosaics in Figure 4.1 are increasing by the complexity of the approaches.

Metadata-based approaches (the position-based and the position- with orientation-based approaches) are susceptible to the sensor errors. These errors appear as a result of the either inaccurate sensors or weak time synchronization between image and sensor. Such errors cause the misalignment in the mosaics. Figure 6.1 shows the position-based mosaicking and Figure 6.2 shows the position- with orientation-based mosaicking of our round mission. Image-based approaches, as presented in Figure 4.1(c), show a good correlation quality. Although this approach usually produces seamless mosaics, it suffers from the problem of error accumulation. In Section 6.2 some results of this approach are shown.

The computation time for the whole set of 37 images in the scaled resolution of $400 \times 300\,px$ took $t_{pos} = 17.31\,s$ for position-based, $t_{pos+rot} = 18.33\,s$ with rota-

	Reference	Pos	Pos + Rot	Image	Hybrid
$\lvert\overline{P_3P_6}\rvert$ [m]	31	31.54	30.53	30.13	31.30
$\lvert\overline{P_6P_{11}}\rvert$ [m]	37.9	38.17	38.07	38.27	38.19
$\lvert\overline{P_3P_{11}}\rvert$ [m]	51.75	50.61	50.76	50.93	52.40
$\lambda_{spat}(O_8)$ [%]		95.3	96.1	94.6	96.9
$\lambda_{corr}(O_8)$ [%]		69.6	74.5	82.4	86.7
$\lambda(O_8)$ [%]		82.4	85.3	88.5	91.8

Table 6.1: Spatial accuracy and quality parameters of the three basic and the hybrid mosaicking approaches. The results are calculated for the first 8 images of the round trip image sequence.

tion, and increased dramatically to $t_{image} = 459.20\,s$ in the image-based alignment approach.

6.1.3 Hybrid approach

We use the same round trip mission to evaluate the hybrid approach (Figure 6.3). As shown we closed the loop of incremental mosaicking (without global optimization) which implies that the mosaicking errors are not accumulated. The computation time was $t_{hybrid} = 136.28\,s$ for the whole set of images, which is significantly less than the image-based approach. The total error range (cp. Figure 4.2) we have used in Figure 6.3 is $GPS_{error} + tan(\alpha) \times \text{height} \simeq 7\,m$ in real world distance at the ground level, which is approximately equivalent to $\frac{1}{4}$ of the image width.

In each image, the only point which has the complete orthogonal view is the nadir point (the point directly under the camera). In a complete nadir-view, this point is the center of the image. In other words, the camera looks to the border of the image with the maximum angle and it only looks orthogonal to the nadir point. It means that the orthogonality will be reduced when getting away from the optical axis. Since we aim to get close to a parallel projection (cp. Section 2.3.2) as much as possible, it gives us an idea that the middle parts of an images contain more reliable data. Therefore, we make sure that the central part of each image under each picture-point is not masked by the border parts of other images. It is done by placing first a central cropped region of each newly added image over the background, and then we place the rest of the image only if the background is empty.

In Figure 6.4, the upper graph shows the relation between correlation of the overlapping parts of two adjacent images in different approaches. As we see the hybrid approach shows the highest correlation as compared to the others; the lower graph indicates the normalized distance between the estimated position and the corresponding GPS position on each image in the hybrid approach (normalization is done by dividing the distances by the image width). By comparing these two graphs, we see that if the estimated position of an image is close to its indicated GPS position it results in a higher correlation and vice versa.

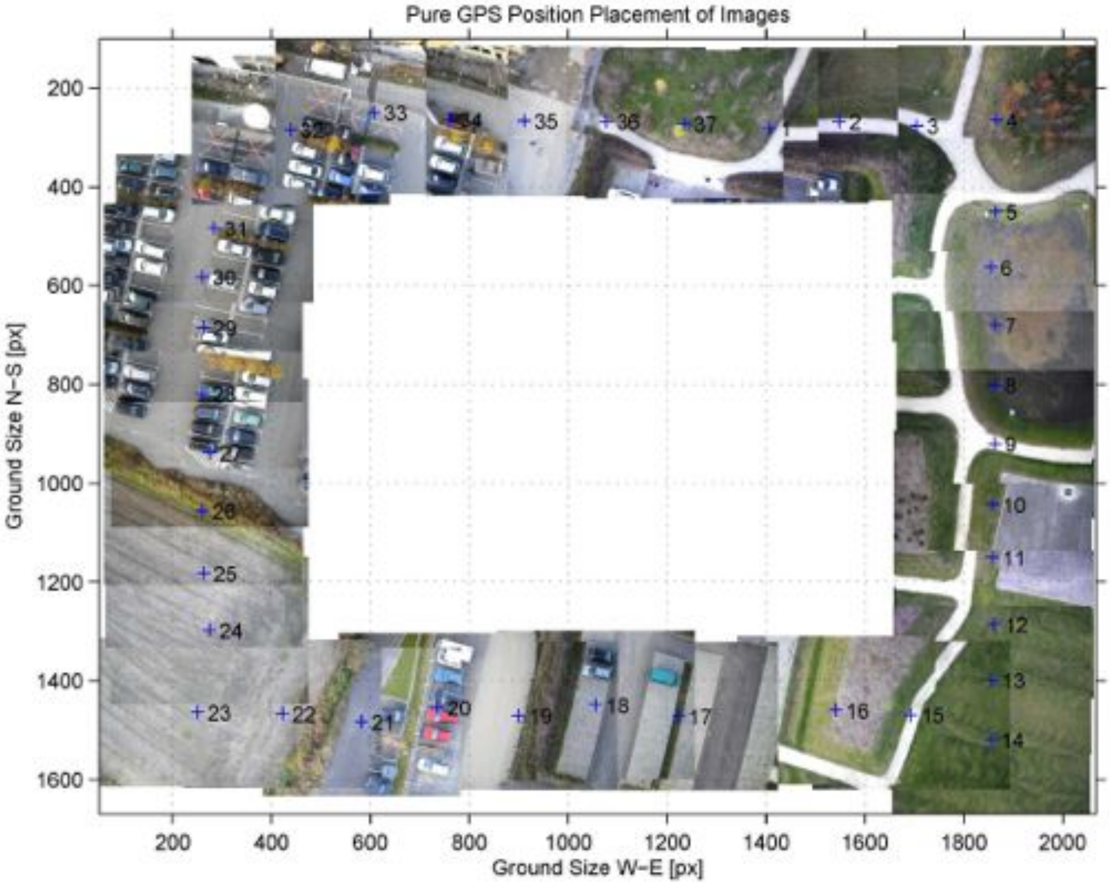


Figure 6.1: Mosaicking result of images taken from a round trip mission using the position-based approach.

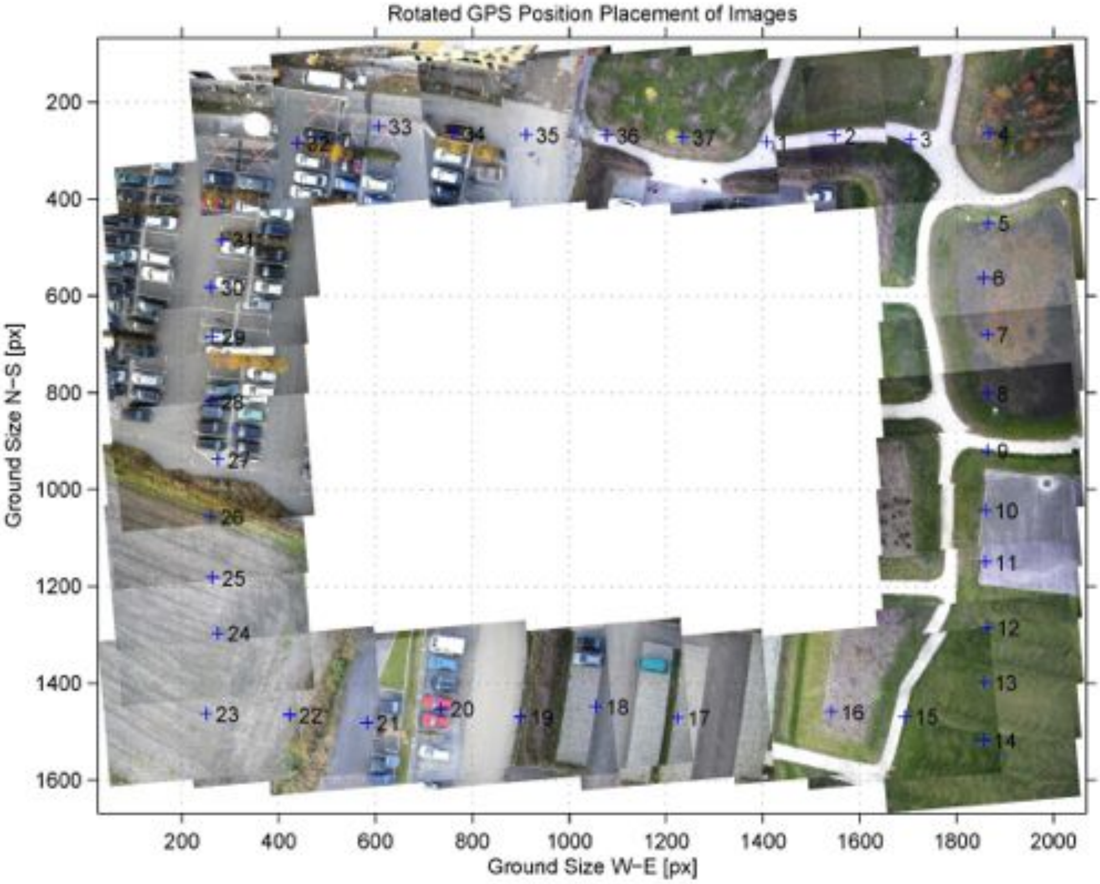


Figure 6.2: Mosaicking result of images taken from a round trip mission using the position- with orientation-based approach.

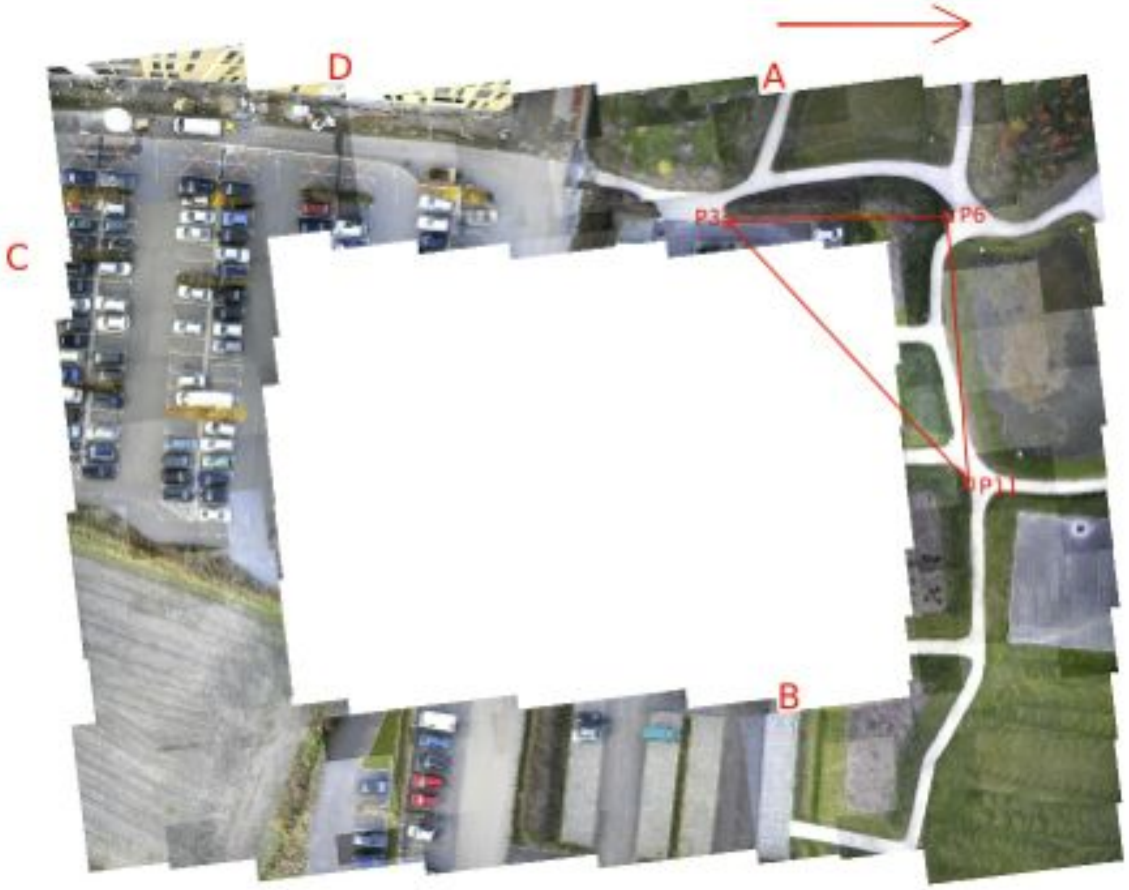


Figure 6.3: Mosaicking result of images taken from a round trip mission using the hybrid approach.

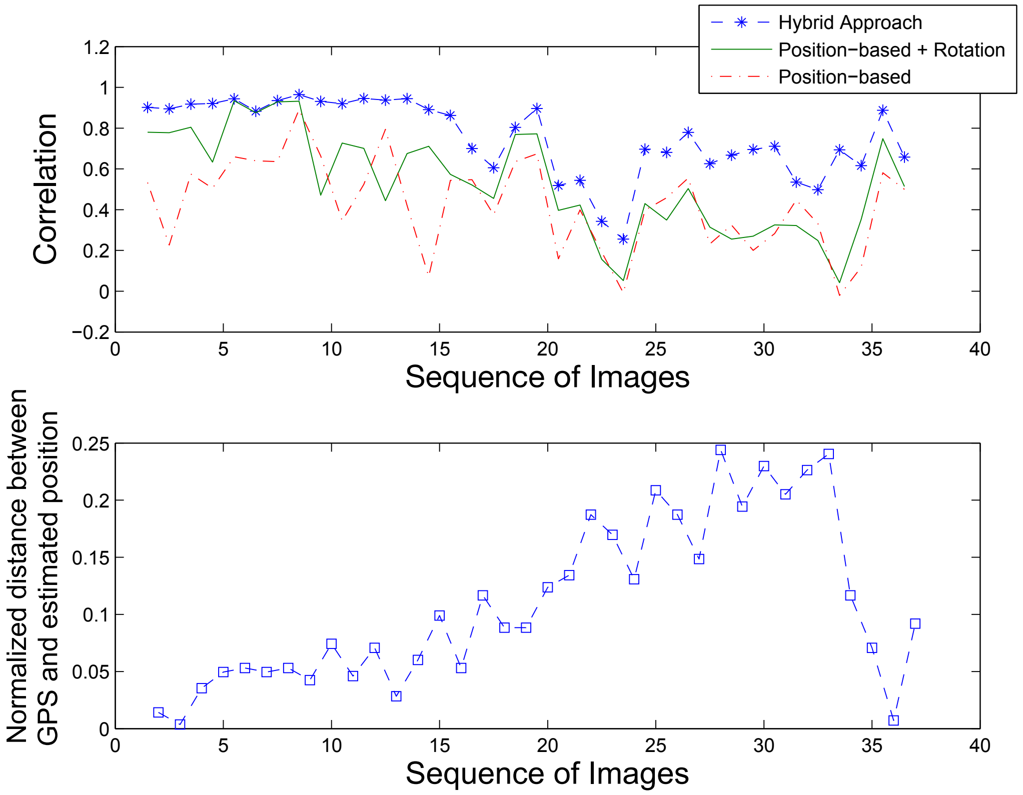


Figure 6.4: The upper graph depicts a comparison between correlation of the overlapping parts of two adjacent images in different approaches; the lower graph shows the normalized distance between the estimated position and the GPS position

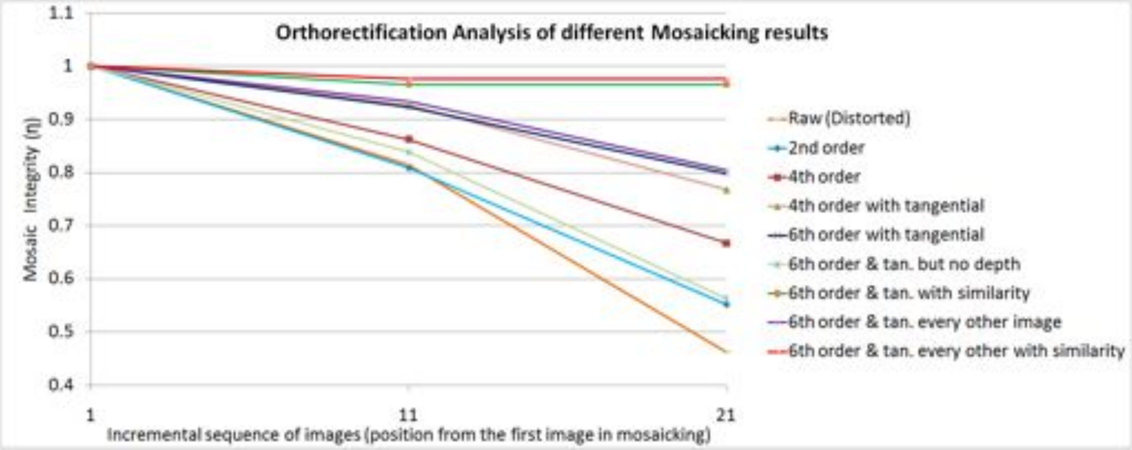
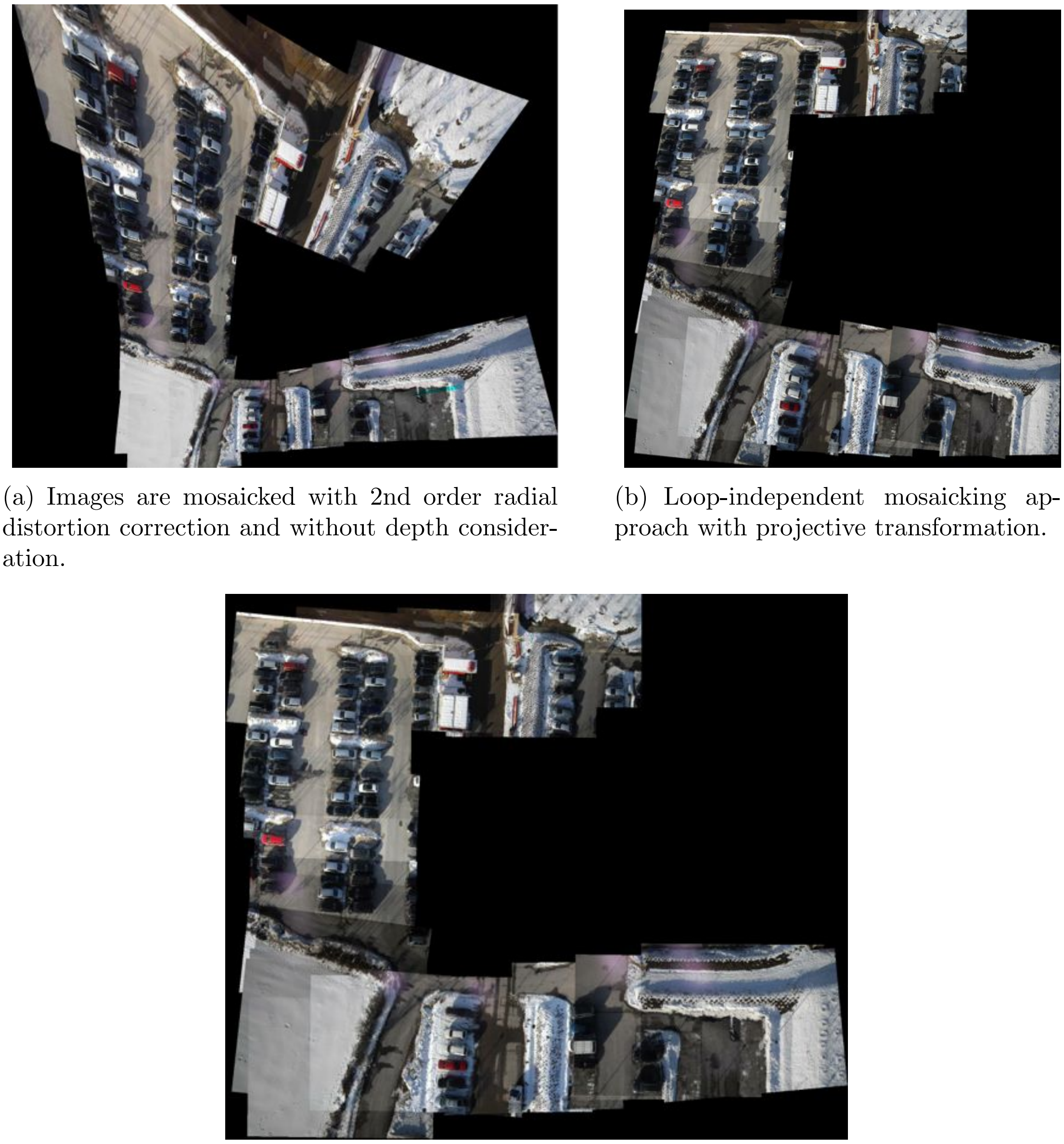


Figure 6.5: Comparison of orthorectification in different mosaics, built with different methods.

6.2 Loop-independent mosaicking

In Section 4.2.1 we lined up a set of printed chess-board patterns to illustrate the loop-independent mosaicking. Here we evaluate the affect of different methods and parameters on orthorectification in such mosaics. In our evaluation we consider the lens distortion correction with different parameters, depth information of the scene, and choice of the transformation model. In Figure 6.5 we use the metric η introduce in Equation 5.1.2 to show how much each mentioned approach or parameter affects the mosaic integrity. As can be seen, using higher orders for radial distortion correction, tangential distortion correction, considering the depth information, and using similarity transformation, all are the factors which can hep us to persist the correct size and preserve the relative distances along the incremental mosaicking process. This affect might not be sensed while using just a couple of images. As shown in this chart, the difference between the 4th order and the 6th order radial distortion correction is not noticeable till the middle of the mosaic, but eventually we can see that 6th order leads to a slightly better quality. It also implies that similarity transformation significantly helps to mitigate the deformation error, since it does not produce and propagate any projective deformation.

Now we show resulting mosaics of images taken by a UAV. In this scenario we took 27 images with approximately 60% of pairwise overlap. Figure 6.6(a) depicts the mosaicking result after 2nd order radial distortion correction without considering the depth information. Figures 6.6(b) and 6.6(c) show the corresponding mosaic considering the loop-independent mosaicking with projective and similarity transformations, respectively. As we expected and as shown in Figure 6.6, mitigating the mentioned errors noticeably improves the orthorectification.



(a) Images are mosaicked with 2nd order radial distortion correction and without depth consideration.

(b) Loop-independent mosaicking approach with projective transformation.

(c) Loop-independent mosaicking approach with similarity transformation.

Figure 6.6: Resulting mosaic of 27 images taken from UAV.

6.3 Thermal lens distortion correction

To evaluate our thermal lens distortion correction method we need to perform a calibration over the outcome of our algorithm, which is a set of the extracted centers of chess-board. To do that, we exploit the well-known MATLAB calibration toolbox developed by Bouguet [4]. Most of the existing calibration methods are designed in a way that they calculate the corners of chess-board pattern as their input. In order to be able to feed the centers of the squares, we have modified the code in a way that it accepts the output of our algorithm which is the centers of the squares. Hence, we calculate the union of the centers of the squares obtained from erosion and dilation (cp. Figure 5.5). The achieved result is shown in Figure 6.7(b). In Figure 6.7(a) the automatic corner extraction result (from Bouguet toolbox) over Figure 5.2(a) is depicted.

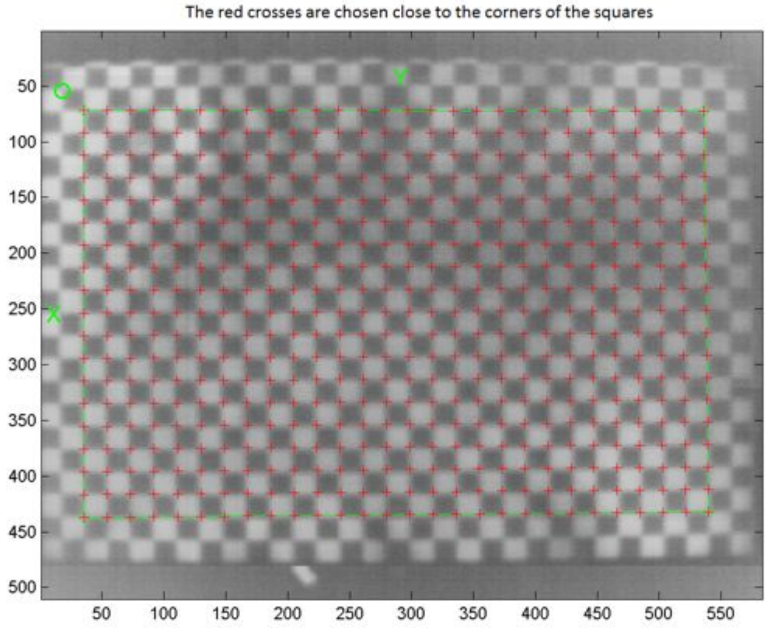
The main goal of our research was to construct a visible pattern for thermal cameras which is possible to perform a calibration over it. This task is fulfilled since the existing calibration toolbox accepts the previously mentioned corner extraction result (cp. Figure 6.7(a)). The corresponding calibration result is shown in Table 6.2 and Figures 6.8(a) and 6.9(a).

The Brown distortion model (cp. Section 2.1.2) can tackle the radial and tangential distortion which is used in our calibration method. But since the error tolerance for $4th$ order of radial distortion and for all orders of tangential distortion are higher than their distortion coefficient so we ignore them. Hence, we only consider the second order of the radial distortion including the principal point estimation which is usually sufficient for narrow fields of view such as our thermal camera. The calibration result performed by our algorithm is depicted in Table 6.3 and Figures 6.8(b) and 6.9(b). By feeding our extracted center of squares to the calibration toolbox with the same parameters we obtain a more accurate result in a sense that the reprojection error depicted in Figures 6.9(a) has lower variance and mean as compared to Figure 6.9(b). As shown in Tables 6.2 and 6.3, the average pixel error $(\mu(X-\hat{X}), \mu(Y-\hat{Y}))$ in our algorithm is $(0.36, 0.38)$ while that of the automatic corner detection method is $(0.49, 0.51)$.

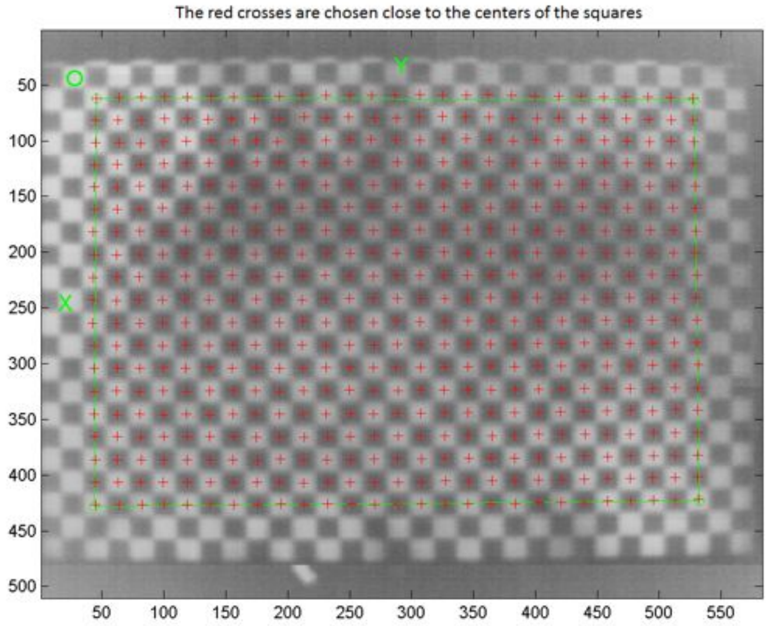
Pixel error	= (0.4872, 0.5059)	
Focal length	= (1297.26, 1402.28)	+/- (559.4, 604.6)
Principal point	= (291, 254.5)	+/- (0, 0)
Skew	= 0	+/- 0
Radial coefficients	= (-0.3906, 0, 0)	+/- (0.3373, 0, 0)
Tangential coefficients	= (0, 0)	+/- (0, 0)

Table 6.2: Intrinsic parameters and the error tolerance. Calibration with automatic corner detection.

Here we show how lens distortion correction affects a real case scenario. In our practical scenario, we aim to register the thermal and visual aerial images taken

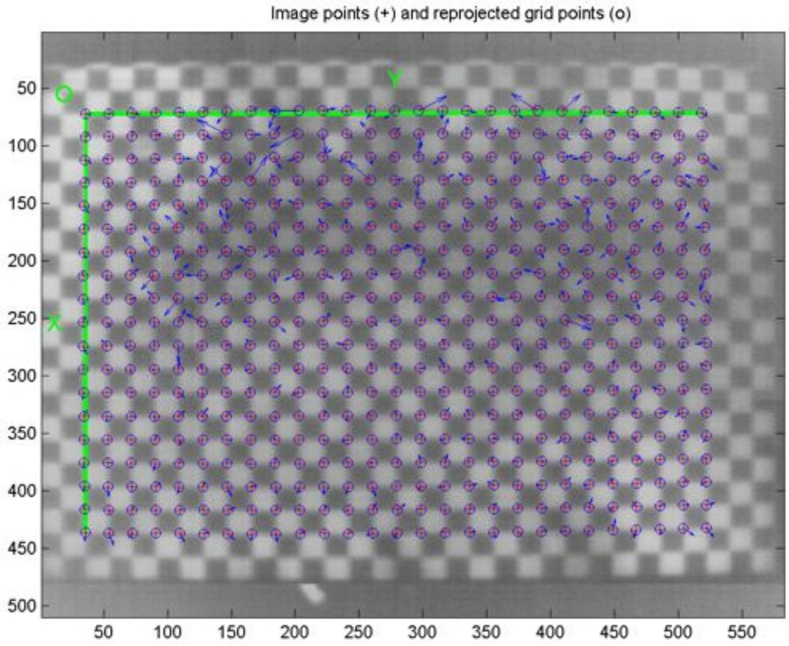


(a) Set of corners extracted by MATLAB calibration toolbox [4].

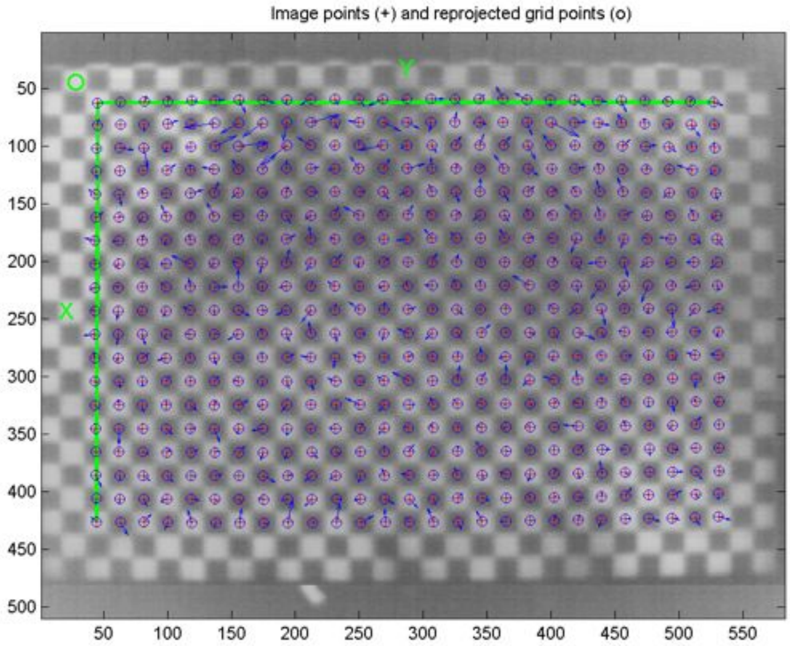


(b) Set of all centers extracted by our algorithm.

Figure 6.7: We extract the centers of squares compared to conventional corner detection.

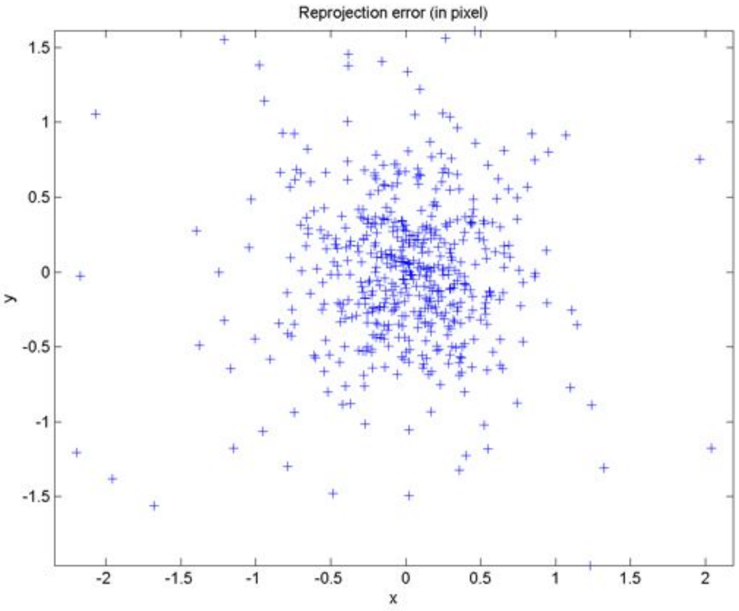


(a) Output of calibration with automatic corner detection.

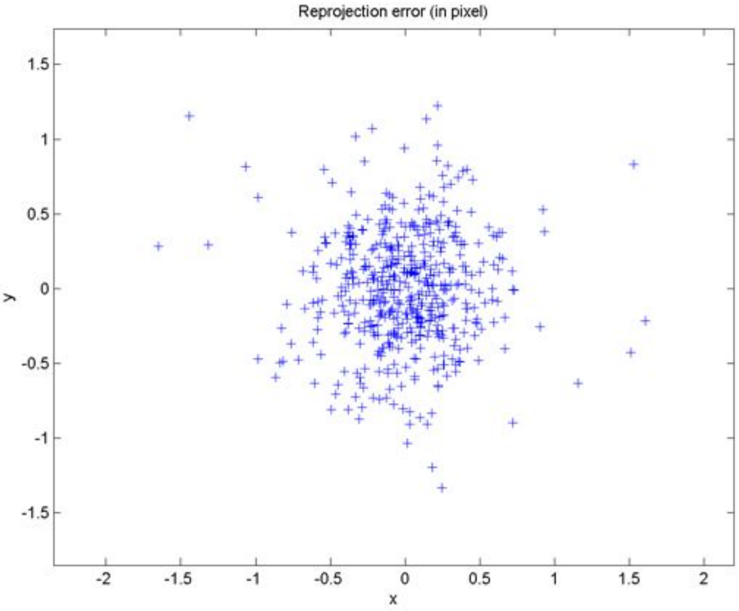


(b) Output of calibration with our center of squares input.

Figure 6.8: Reprojection error over the pattern.



(a) Output of calibration with automatic corner detection.



(b) Output of calibration with our center of squares input.

Figure 6.9: Chart of reprojection error.

Pixel error	= (0.3599, 0.3845)	
Focal length	= (1041.67, 1125.61)	+/- (491.1, 530.6)
Principal point	= (291, 254.5)	+/- (0, 0)
Skew	= 0	+/- 0
Radial coefficients	= (-0.2698, 0, 0)	+/- (0.2548, 0, 0)
Tangential coefficients	= (0, 0)	+/- (0, 0)

Table 6.3: Intrinsic parameters and the error tolerance. Calibration with our center of squares input.

from the same scene by exploiting the UAVs shown in Figure 1.1. We made a test set of some pairs of thermal and visual aerial images. This test set consists of both distorted and undistorted aerial images. Hence, over each pair of images we extract the matching feature points by using SURF feature extraction method. Afterwards registration procedure is performed once over the raw (distorted) images and once over the undistorted images (cp. Figure 6.10). Green lines in Figures 6.10(a) and 6.10(b) imply the corresponding feature points. In Figures 6.10(c) and 6.10(d) the registered result after homography is shown.

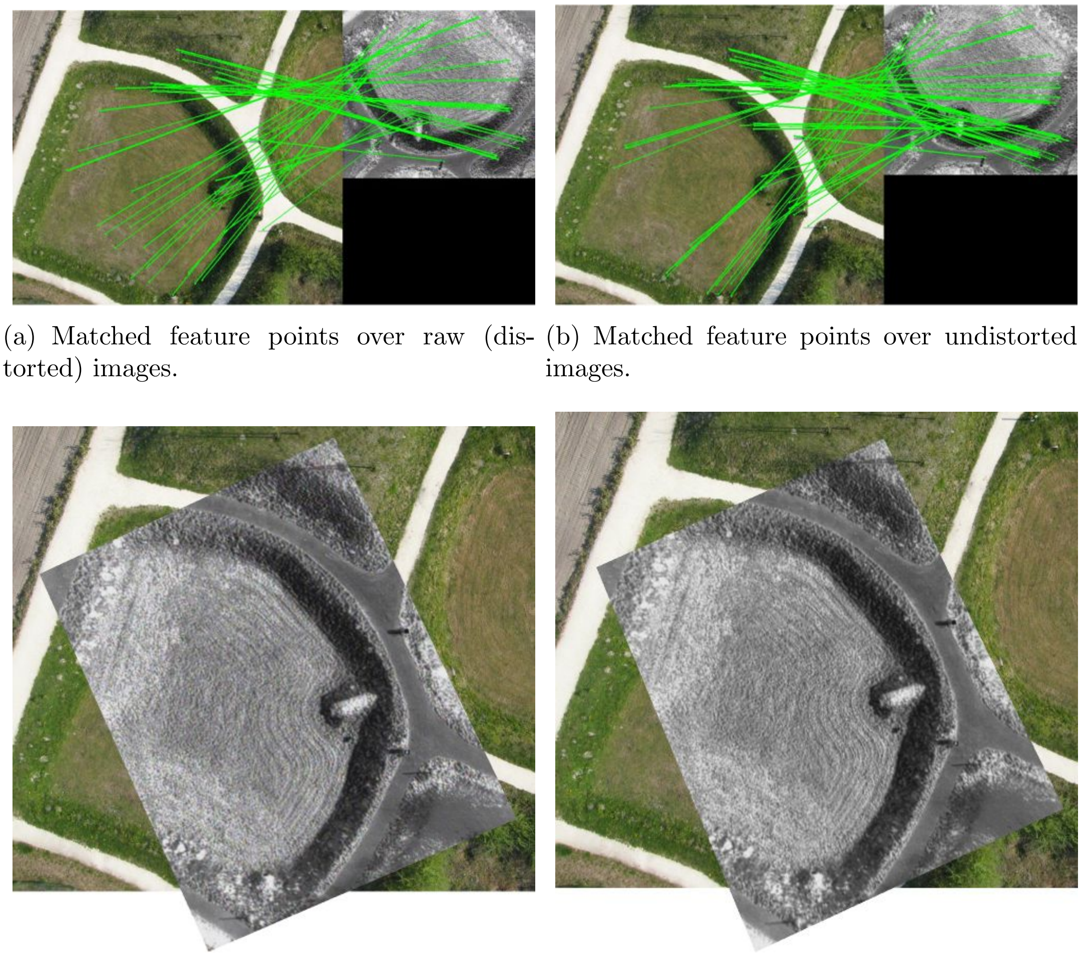
To be able to quantify the improvement of the results after distortion correction, we use the mean squared error (MSE) as our error function:

$$MSE(\hat{P}) = \frac{\sum_{i=1}^{n} (\hat{p}_i - p_i)^2}{n} \tag{6.2}$$

where $P = \{p_i | i = 1 \ldots n\}$ is the set of the feature points in the reference image, $\hat{P}$ is the estimation of P which is obtained by performing the homography over the corresponding feature points of the transformed image and n is the number of feature points. This error for raw images in Figure 6.10(a) with 38 feature points is $MSE = 13.2$, and for undistorted images in Figure 6.10(b) with 54 feature points is $MSE = 11.7$. We have calculated this error with 5 different sets of feature points over 4 pairs of images and the statistics shows that the MSE error when registering the undistorted images is reduced in average by 17% as compared to raw image registration. Obviously using a wider lens produces a higher barrel distortion and consequently in that case the mean squared error will show a higher reduction after distortion correction.

6.4 Evaluation of interspectral registration

This section presents further experimental results and discussions regarding interspectral registration. First, we evaluate the performance of the RFAE wrt. other feature extraction methods. Second, we extend the discussion on our registration by exploiting images mosaics and depth maps.



(a) Matched feature points over raw (distorted) images.

(b) Matched feature points over undistorted images.

(c) Registered after homography of raw (distorted) images.

(d) Registered after homography of undistorted images.

Figure 6.10: Registration of thermal and visual aerial images.

6.4.1 RFAE evaluation

The evaluation of the RFAE methods is based on a heterogeneous data-set of 84 image pairs of different spectrums. This data-set consists of different types of satellite images, images of human bodies, general surveillance images, and aerial images from low-altitude UAVs. The resolution varies for visual images between $320 \times 240\,px$ and $1047 \times 1061\,px$ and for thermal images between $320 \times 240\,px$ and $584 \times 512\,px$. The overlap ratio between the image pairs varies between 50% and 100%.

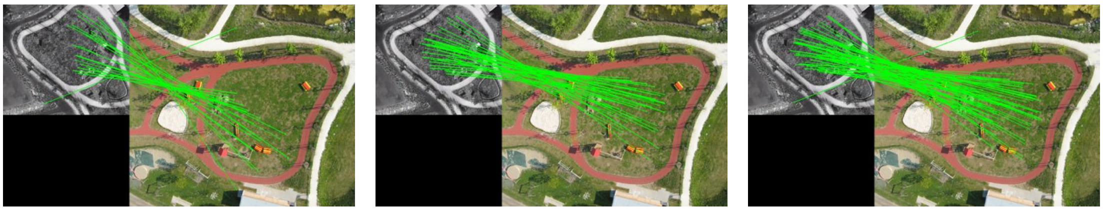
We performed the interspectral image registration over this data-set by using the SIFT, SURF, upright SURF, RFAE, and combination of SURF with RFAE. We use a quality metric to evaluate the extracted features for the purpose of interspectral registration. Although the number of corresponding matched features (inliers) is often used to evaluate a registration, it does not provide any information regarding the distribution of the features. In our case, we use this number for acceptance or rejection of a registration based on a threshold. If a registration is accepted, we can use our quality metric to evaluate it or compare it with other methods. The success level of a registration increases when there are sufficient inliers and they are distributed uniformly over the image. Nevertheless, a metric which is modeling the deviation from a uniform distribution (cp. [54]) is not appropriate for our case. Such a metric is built to quantify the inhomogeneity, so that adding an additional point very close to (or almost over) an existing point reduces the magnitude of the metric. In our case the metric should not change if we add a point exactly over an existing point and should increase slightly if we add a point in the close neighborhood of an existing point.

Suppose that an optimal feature distribution for the purpose of registration demands at least one feature point within a distance δ to any random point in the image, we construct our metric as follows:

$$Q = \frac{|\bigcup_i \{\mathbf{x} : \|\mathbf{x} - \mathbf{f}_i\| < \delta\}|}{A} \times \max_{i,j} \frac{\|\mathbf{f}_i - \mathbf{f}_j\|}{d}, \tag{6.3}$$

where $\mathbf{x}$ represents a point in the image, $\mathbf{f}$ an inlier feature point in the image, i, j indexes of the inliers, A the area of the image which is equivalent to the image resolution in pixels and d is the length of the diagonal of the image. The denominators of the fractions aim to normalize the metric to the range $[0, 1]$. The numerator of the first fraction represents the aggregated area of all circles with radius δ centered at feature points. The first fraction of the equation implies the coverage ratio of the feature points within the image. The remainder of the equation describes the normalized maximum distance between all possible feature point pairs. This favors the sparse features rather than dense features which is an important factor for a successful registration. The Q value takes its maximum, 100%, if and only if we have at least two features over the two far corners of the image and there exist no circle with radius δ in the image so that no feature falls inside this circle.

Defining the value for δ depends on different factors such as image quality and

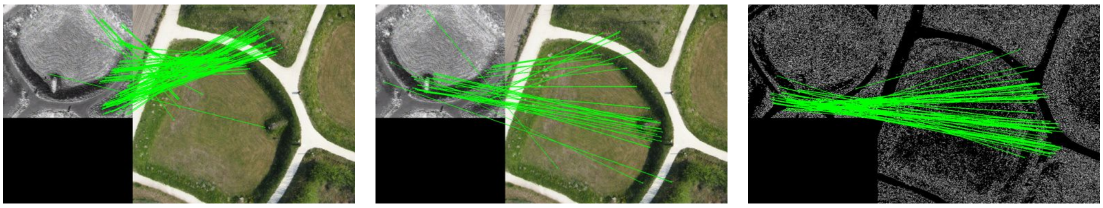


(a) SURF initial lobe=5, inliers=22, Q=17.2%.

(b) SURF initial lobe=3, inliers=38, Q=8%.

(c) SURF initial lobe=1, inliers=62, Q=16.6%.

Figure 6.11: Finding the initial lobe parameter of SURF which maximizes the Q value.



(a) SIFT method with inliers=102 and Q=22%.

(b) SURF method with inliers=31 and Q=23%.

(c) RFAE method with inliers=59 and Q=12%.

Figure 6.12: Finding the best method which maximizes the Q value. In this example SURF has a higher Q value.

resolution. In our experiments we set δ to 10% of the image width. To accept a registration we require at least 9 corresponding feature pairs (inliers). We set the Q value to zero for unaccepted registrations. We use this metric primarily for comparing different methods of registration over the same pair of images. For instance we can use it to find the best parameters for an individual feature extraction method (cp. Figure 6.11) or compare the registration performance between different methods (cp. Figure 6.12). These samples also show that more inliers do not necessarily result in an increase of the Q value. Note that the high Q value in Figure 6.11(a) is caused by a single distant feature point, despite its small number of inliers. The sparse feature distribution in Figure 6.12(b) achieves also a high Q value with far less inliers than in Figure 6.12(a).

This metric is used to identify the best registration for each pair of images and classify our data-set as shown in Table 6.4. The fractions shown in this table represent the ratio of the number of acceptable registrations to all number of pairs.

Since satellite images have a relatively high overlap and are usually aligned quite well with a fixed rotation and scale, most of the feature extraction methods succeed to register these images. In cases with high deviation between the spectral bands (in which SURF, SIFT, and upright SURF failed with the registration), the

	SIFT	SURF	U-SURF	RFAE	SURF+ RFAE
Satellite (low deviation)	24/24	24/24	24/24	24/24	24/24
Satellite (high deviation)	7/10	6/10	9/10	10/10	10/10
Human	0/16	0/16	2/16	12/16	12/16
Surveillance	1/14	2/14	4/14	12/14	12/14
UAV	13/20	14/20	2/20	13/20	17/20

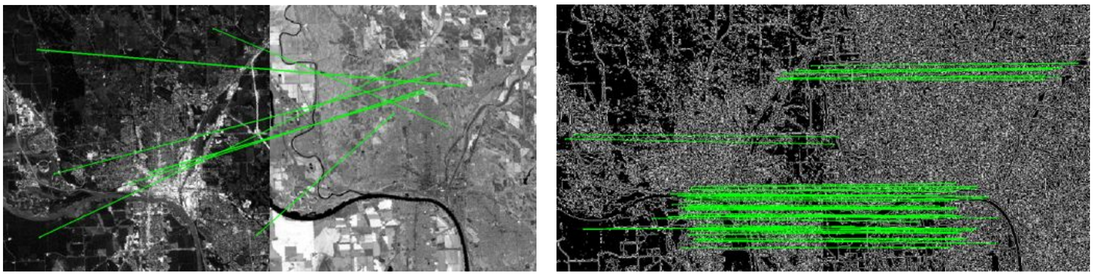
Table 6.4: Successful registration ratios based on different feature extraction methods and types of images.

	SIFT	SURF	U-SURF	RFAE	SURF+ RFAE
Satellite (low deviation)	72%	66%	71%	53%	68%
Satellite (high deviation)	45%	41%	49%	46%	52%
Human	0%	0%	1%	8%	8%
Surveillance	1%	3%	4%	18%	19%
UAV	9%	10%	1%	6%	11%

Table 6.5: Average Q values based on different feature extraction methods and types of images.

RFAE method performs better for registration as shown in Figure 6.13. The RFAE method shows the highest improvement for images of human bodies. Figures 6.14 and 6.15 depict samples of such a thermal and visual image registration. A similar improvement can be seen for surveillance scenarios (cp. Figures 6.16 and 6.17). The interspectral registration of low-altitude aerial images has turned out to be more challenging. Whenever a pair of aerial images does not share enough mutual edge information, the RFAE method shows a weak performance. However, in cases with more mutual edge information (such as Figures 6.18 and 6.19) RFAE dominated the other methods. We therefore combine the RFAE and SURF methods and choose the feature extraction method with highest Q value for the registration of the images. As shown in the last column of Table 6.4, this combination chooses the best result among RFAE and SURF and achieves the best overall registration performance.

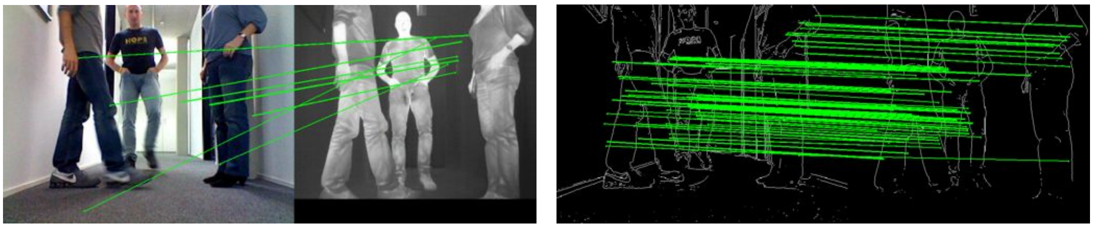
Table 6.5 shows the average Q values of the same data-set used in Table 6.4. Image pairs with high correlation (e.g., most of the satellite images) show a high Q value. For these images, the success rate of registration by using standard methods is higher rather than RFAE. The reason is that standard methods are able to extract more detailed features as compared to RFAE which extracts merely features along the edges. On the other hand, RFAE is dominant when images have a high deviation, yet with sufficient mutual edges. The average Q values corresponding to images of surveillance or human bodies imply the better performance of RFAE. Table 6.4 represents mainly the interspectral registration acceptance rate while Ta-



(a) Failed registration using SURF.

(b) Successful registration using RFAE.

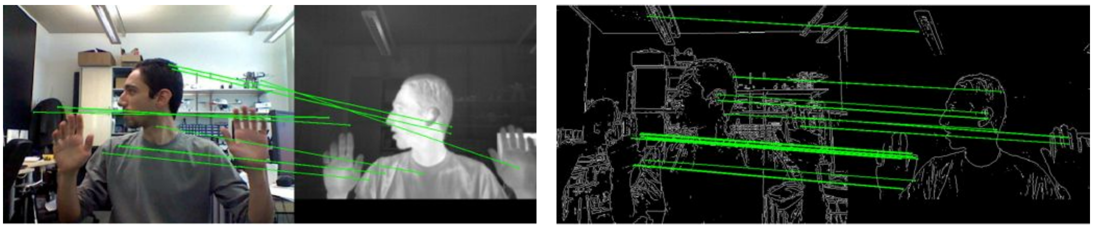
Figure 6.13: Registration between bands 1 and 4 of the Landsat satellite image of Iowa state (image source: NASA/USGS).



(a) Failed registration using SURF.

(b) Successful registration using RFAE.

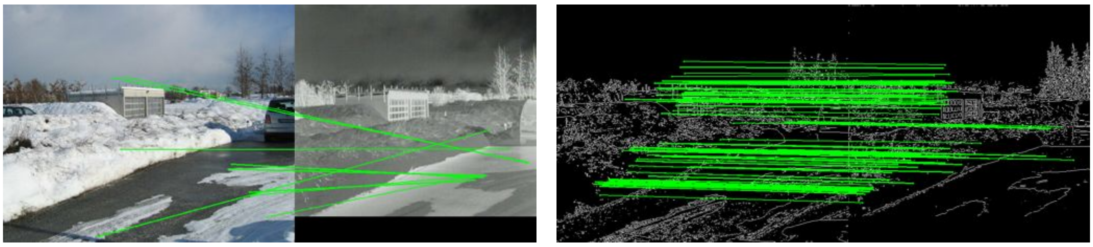
Figure 6.14: Registration of thermal and visual images of humans.



(a) Failed registration using SURF.

(b) Successful registration using RFAE.

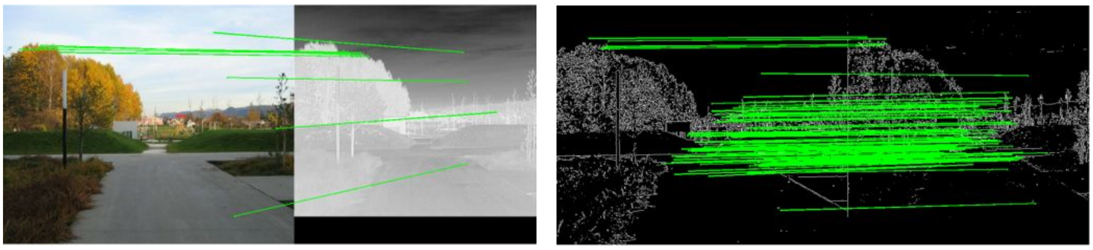
Figure 6.15: Registration of thermal and visual images of humans.



(a) Failed registration using SURF.

(b) Successful registration using RFAE.

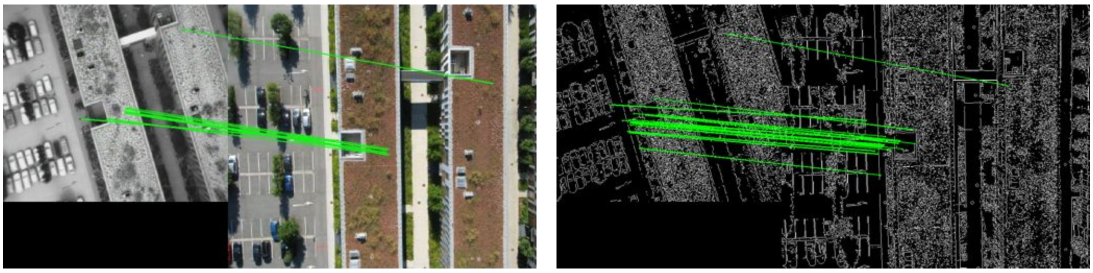
Figure 6.16: Registration of thermal and visual surveillance images.



(a) Failed registration using SURF.

(b) Successful registration using RFAE.

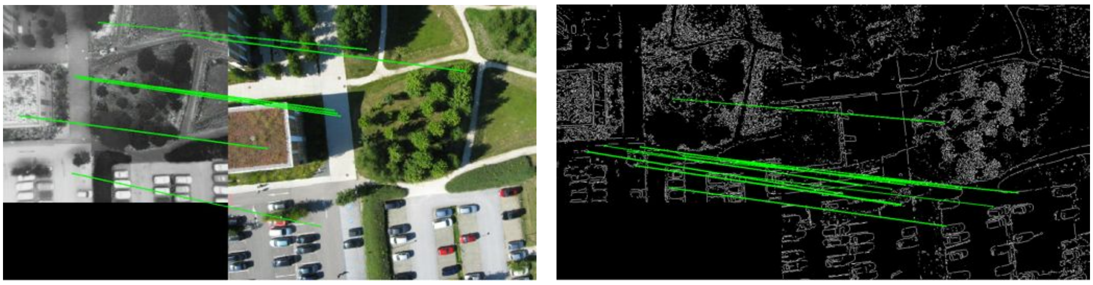
Figure 6.17: Registration of thermal and visual surveillance images.



(a) Failed registration using SURF.

(b) Successful registration using RFAE.

Figure 6.18: Registration of thermal and visual aerial images taken from low-altitude UAV.



(a) Failed registration using SURF.

(b) Successful registration using RFAE.

Figure 6.19: Registration of thermal and visual aerial images taken from low-altitude UAV.

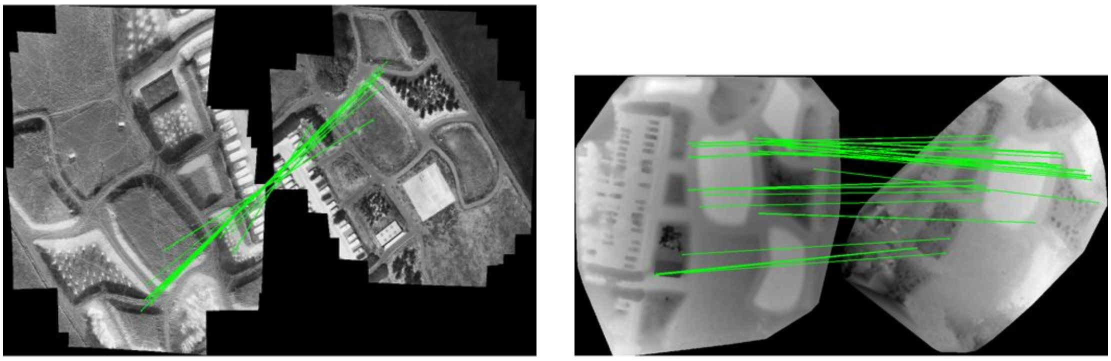
ble 6.5 represents the quality of extracted features for the purpose of interspectral registration.

6.4.2 Mosaic registration

As explained in Section 5.2.3, we are able to register two mosaics together as long as there is one pair of images which can be registered among the entire mosaics. Although we can achieve a higher accuracy when more image pairs are registered, there is a general drawback in this type of registration. In most of the constructed mosaics there is some deformation. This can be due to the different angles of imaging from non-flat objects or some non-rigid transformations performed in mosaicking. This problem in addition to the accumulated error near the borders sometime cause some misalignment and ghosting effect in the fusion as you can see in Figure 5.11(b).

6.4.3 Depth map registration

The registration of the depth map instead of the mutual image information is often more complex and computationally expensive (see Section 6.5 for more details). However, there is an advantage in cases of aerial imagery with low temporal resolution. To demonstrate this advantage we tested the registration of aerial images taken in summer and winter (cp. Figure 5.11). The images taken in different seasons exhibit a very high variation. As shown in Figure 6.20(a), images of the same spectrum (i.e., thermal mosaic in winter and thermal mosaic in summer) can be only weakly registered by standard SURF. On the other hand, by our depth map method we were able to successfully register even the more complex scenario of interspectral registration with a high image variation over time. Figure 6.20(b) shows such successful registration between thermal winter mosaic and the visual summer mosaic.



(a) Weak registration between thermal mosaics in summer and winter (same spectrum).

(b) Registration between thermal depth map in winter and visual depth map in summer (interspectral).

Figure 6.20: Registration between two set of aerial images taken in different seasons.

6.5 Further discussion

Although the methods and results for aerial image mosaicking presented in this thesis are distinct and independent, it is possible to combine them together. The steps explained in loop-independent mosaicking are applicable to all mosaicking methods, since mitigating the sources of error is always recommended. It can be combined with the hybrid approach or the multispectral mosaicking as a preliminary step. Note that both mosaicking approaches we have presented here aim to provide orthorectified mosaics, while the first approach (cp. Section 4.1) focus on georeferencing and the second approach (cp. Section 4.2) focus on orthorectification in loop free images. The evaluations show that our approach results in a higher correlation between overlapping image regions and retains spatial distances with an error of less than $30\,cm$. The computation time for a set of 37 images is reduced by approximately 70% compared to an image-based mosaicking.

In interspectral registration methods and multispectral mosaic construction we have mainly exploited our own mosaicking methods. The independence of these methods makes it easier to combine them together. Figures 5.10 and 5.11 show sample mosaics in which we used only the methods explained in this thesis. The computational complexity of the interspectral registration between two mosaics (considering each mosaic from n individual images) is approximately n times more than a single-pair registration. Using Matlab on a standard PC running at $2.66\,GHz$, we perform on average a single-pair registration in $2\,s$, while the registration of the mosaics in Figure 5.11 is performed in $57\,s$.

The computation time for the depth map registration based on disparity vectors (cp. Figure 5.12) is on average 4 times more than an equivalent single-pair registration. This is because to construct a pair of depth maps first we need to register two pairs of images. Therefore, in total we perform 3 registrations and 2 depth map constructions. However, the computational complexity of the depth map registration based on a full bundle adjustment is much higher. For sample registration shown in Figure 5.14, the computation time is about 2 hours.

CHAPTER

7 Conclusions and future work

In this thesis we presented our system for mosaicking high-resolution overview images of large areas with high geometric accuracy from a set of images taken from small-scale UAVs. Although much research has been done on mosaicking of aerial imagery, the challenges in our application are significantly different since small-scale UAVs flying at low altitude pose new problems. We propose a hybrid approach that combines inaccurate information on the camera's position and orientation, and the image data. Thus, we can maintain geometric accuracy and at the same time enhance the visual appearance. The evaluations show that the hybrid approach results in a higher correlation between overlapping image regions and also preserves the relative distances. The computation time for a set of 37 images is reduced by approximately 70% compared to an image-based mosaicking.

We also quantify the influence of different parameters such as sensor distortion model, depth information of the scene, and the choice of projection and transformation models over sequential, pairwise and loop-free image mosaicking. Understanding and comparing the sources of errors enables us to minimize those errors in a way that increases the orthorectification in aerial image mosaicking. Using higher polynomial orders in geometric distortion correction might not be noticeable in a pair of images, but at some point in incremental image mosaicking it will show its affect. To retain the relative distances, similarity transformation, despite its lower degree of freedom, is a good substitution for projective transformation if we have almost a nadir-view of the camera. It is also shown how a simple depth map help us to choose the appropriate feature points on the ground plane for an accurate mosaicking.

Further, we have shown how to perform a robust interspectral image registration. In general we proposed some methods to register two images sensed in different spectrums, however we mainly focused on thermal and visual image registration. First we presented a general method (RFAE) which exploits the existing scale-invariant feature extraction methods such as SIFT and SURF in order to extract the robust features along the edges. Based on experimental result, our approach have shown a noticeable improvement in interspectral registration. Second we proposed two methods for increasing the robustness and extracting additional features in cases with more than one pair of images. The latter scenario was studied with a focus on the thermal and visual aerial images taken from low-altitude UAVs. In case of multiple image pairs, we showed either we can use the image mosaics for the interspectral registration or we can use depth maps of a target scene for the feature extraction.

The future work and open issues in this field vary based on the approach:

- Different types of UAV may pose different challenges. With advances in technology and with higher processing power it is possible to process a larger number of images with higher resolution. Another possible scenario is to consider aerial videos for mosaicking.

- In our hybrid approach the combination of two transformations from image-based and metadata-based methods is performed by simulated annealing method. The combination method can be studied and can be improved by comparing to other heuristic methods.

- In our loop-independent mosaicking we have considered the scenarios with no loop in the image sequence. One can extend our method of orthorectified mosaicking to combine with a heuristic method of global optimization in presence of loop.

- In the process of interspectral registration, more complex features such as classified areas or object can be used. However this will increase the computational complexity of the feature extraction and matching.

- At the system level, future works may focus on underlying architectures for deployment of multi-UAV for the purpose of aerial imagery. This includes the autonomous planning and deployment, communication structure, and optimal coverage methods.

Bibliography

[1] A. Agarwala, M. Agrawala, M. Cohen, D. Salesin, and R. Szeliski. Photographing long scenes with multi-viewpoint panoramas. *ACM Transactions on Graphics (TOG)*, 25:853–861, July 2006.

[2] P. Azzari, L. Stefano, and S. Mattoccia. An evaluation methodology for image mosaicing algorithms. In *Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 89–100, Berlin, Heidelberg, 2008. Springer-Verlag.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[4] J.-Y. Bouguet. Camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/, February 2013.

[5] A. Brown, C. Gilbert, H. Holland, and Y. Lu. Near real-time dissemination of geo-referenced imagery by an enterprise server. In *Proceedings of Canadian Cartographic Association Conference - GeoTec Event*, Ottawa, Ontario, Canada, June 2006.

[6] D. C. Brown. Decentering distortion of lenses. *Photogrammetric Engineering*, 32(3):444–462, 1966.

[7] H. Burdick. *Digital Imaging: Theory and Applications*. McGraw-Hill, 1997.

[8] C. Çiğla and A. A. Alatan. Multi-view dense depth map estimation. In *Proceedings of the 2nd International Conference on Immersive Telecommunications (IMMERSCOM)*, pages 1–6, ICST, Brussels, Belgium, 2009. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[9] E. Corias, J. Santamaria, and C. Miravet. A Segment-based Registration Technique for Visual-IR Images. *Optical Engineering*, (1):1–29, 2000.

[10] Q. Du and N. Raksuntorn. Automatic registration and mosaicking for airborne multispectral image sequences. *Photogrammetric Engineering & Remote Sensing*, 74(February):169–181, 2008.

[11] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 612–618, 2000.

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

[13] S. Flöry and M. Hofer. Surface fitting and registration of point clouds using approximations of the unsigned distance function. *Computer Aided Geometric Design*, 27(1):60–77, January 2010.

[14] L. M. G. Fonseca and M. H. M. Costa. Automatic registration of satellite images. In *Proceedings of Brazilian Symposium on Computer Graphics and Image Processing*, pages 219–226, 1997.

[15] Y. Furukawa and J. Ponce. Accurate camera calibration from multi-view stereo and bundle adjustment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number 3, pages 1–8, Hingham, MA, USA, 2008.

[16] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, August 2010.

[17] G. Gaussorgues. *Infrared Thermography.* Microwave Technology Series. Chapman & Hall, 1994.

[18] X. Han, H. Zhao, L. Yan, and S. Yang. An approach of fast mosaic for serial remote sensing images from uav. In *Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 11–15, Washington, DC, USA, 2007. IEEE Computer Society.

[19] B. B. Hansen and B. S. Morse. Multiscale image registration using scale trace correlation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 202–208, 1999.

[20] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of Fourth Alvey Vision Conference*, pages 147–151, 1988.

[21] G. Hong and Y. Zhang. The image registration technique for high resolution remote sensing image in hilly area. In *Proceedings of International Society of Photogrammetry and Remote Sensing (ISPRS) joint conference*, 2005.

[22] R. Hruska, G. Lancaster, J. Harbour, and S. Cherry. Small uav-acquired, high-resolution, georeferenced still imagery. In *Proceedings of AUVSI Unmanned Systems North America*, pages 837–840, June 2005.

[23] Y. Huang, J. Li, and N. Fan. Image mosaicing for UAV application. In *Proceedings of the International Symposium on Knowledge Acquisition and Modeling (KAM)*, pages 663–667, Washington, DC, USA, 2008. IEEE Computer Society.

[24] R. Istenic, D. Heric, S. Ribaric, and D. Zazula. Thermal and visual image registration in Hough parameter space. In *Proceedings of 14th International Workshop on Systems, Signals and Image Processing (IWSSIP)*, pages 106–109, 2007.

[25] A. Jensen, M. Baumann, and Y. Chen. Low-cost multispectral aerial imaging using autonomous runway-free small flying wing vehicles. *Geoscience and Remote Sensing Symposium, IGARSS*, 5:506–509, 2008.

[26] J. W. Joo, J. W. Choi, and D. L. Cho. Robust registration in two heterogeneous sequence images on moving objects. In *Proceedings of Sixth International Conference of Information Fusion*, pages 277–282, 2003.

[27] S. B. Kang and R. Szeliski. Extracting view-dependent depth maps from a collection of images. *International Journal of Computer Vision*, 58(2):139–163, 2004.

[28] J. P. Kern and M. S. Pattichis. Robust multispectral image registration using mutual-information models. *IEEE Transactions on Geoscience and Remote Sensing*, 45(5):1494–1505, 2007.

[29] H. Kim and M.-G. Kim. Image registration using terrain relief correction based on the rigorous sensor models. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XXXIX-B1, pages 235–238, 2012.

[30] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[31] S. G. Kong, J. Heo, F. Boughorbel, Y. Zheng, B. R. Abidi, A. Koschan, M. Yi, and M. A. Abidi. Multiscale fusion of visible and thermal IR images for illumination-invariant face recognition. *International Journal of Computer Vision*, 71(2):215–233, June 2006.

[32] G. B. Ladd, A. Nagchaudhuri, M. Mitra, T. J. Earl, and G. L. Bland. Rectification, georeferencing, and mosaicking of images acquired with remotely operated aerial platforms. In *Proceedings of American Society for Photogrammetry and Remote Sensing (ASPRS)*, page 10 pp., Reno, NV, USA, May 2006.

[33] S. R. Lee. A coarse-to-fine approach for remote-sensing image registration based on a local method. *International Journal on Smart Sensing and Intelligent Systems*, 3(4):690–702, 2010.

[34] H. Lin, P. Du, W. Zhao, L. Zhang, and H. Sun. Image registration based on corner detection and affine transformation. In *Proceedings of 3rd International Congress on Image and Signal Processing*, pages 2184–2188, October 2010.

[35] L. Lou, F.-M. Zhang, C. Xu, F. Li, and M.-G. Xue. Automatic registration of aerial image series using geometric invariance. In *Proceedings of IEEE International Conference on Automation and Logistics*, pages 1198–1203, 2008.

[36] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[37] H. Mahdi and A. A. Farag. Image registration in multispectral data sets. In *Proceedings of IEEE International Conference on Image Processing*, volume 2, pages 369–372, 2002.

[38] C. Micheloni and G. L. Foresti. Real-time image processing for active monitoring of wide areas. *International Journal of Visual Communication and Image Representation (JVCIR)*, 17(3):589–604, 2006.

[39] C. Micheloni and G. L. Foresti. Active tuning of intrinsic camera parameters. *IEEE Ttransactions on Automation Science and Engineering*, 6(4):577–587, October 2009.

[40] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, October 2004.

[41] J.-M. Morel and G. Yu. Is SIFT scale invariant? *Inverse Problems and Imaging (IPI)*, 5(1):115–136, 2011.

[42] P. Pesti, J. Elson, J. Howell, D. Steedly, and M. Uyttendaele. Low-cost orthographic imagery. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems (GIS)*, pages 1–8, New York, NY, USA, 2008.

[43] M. Quaritsch, K. Kruggl, D. Wischounig-Strucl, S. Bhattacharya, M. Shah, and B. Rinner. Networked UAVs as aerial sensor network for disaster management applications. *e&i Journal*, 127(3):56–63, March 2010.

[44] M. Quaritsch, R. Kuschnig, H. Hellwagner, and B. Rinner. Fast aerial image acquisition and mosaicking for emergency response operations by collaborative UAVs. In *Proceedings of 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2011)*, Lisbon, Portugal, 2011.

[45] M. Quaritsch, E. Stojanovski, C. Bettstetter, G. Friedrich, H. Hellwagner, B. Rinner, M. Hofbauer, and M. Shah. Collaborative microdrones: Applications and research challenges. In *Proceedings of the Second International Conference on Autonomic Computing and Communication Systems*, 2008.

[46] M. Quaritsch, D. Wischounig-Strucl, S. Yahyanejad, V. Mersheeva, E. Yanmaz, G. Friedrich, H. Hellwagner, C. Bettstetter, and B. Rinner. Collaborative microdrones research questions & challenges. In *Proceedings of International Workshop on Self-Organizing Systems*, page 38. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011.

[47] M. Quaritsch, D. Wischounig-Strucl, S. Yahyanejad, V. Mersheeva, E. Yanmaz, G. Friedrich, H. Hellwagner, C. Bettstetter, and B. Rinner. FAMUOS: A multi-UAV system for aerial reconnaissance in rescue scenarios. In *Proceedings of Austrian Robotics Workshop*, 2011.

[48] B. Rinner, M. Quaritsch, D. Wischounig-Strucl, and S. Yahyanejad. Apparatus and method for generating an overview image of a plurality of images using a reference plane. European Patent No. EP2423873 (A1), 2012.

[49] B. Rinner, M. Quaritsch, D. Wischounig-Strucl, and S. Yahyanejad. Apparatus and method for generating an overview image of a plurality of images using an accuracy information. European Patent No. EP2423871 (A1), 2012.

[50] J. Roßmann and M. Rast. High-detail local aerial imaging using autonomous drones. In *Proceedings of 12th AGILE International Conference on Geographic Information Science: Advances in GIScience*, Hannover, Germany, June 2009.

[51] P. Rudol and P. Doherty. Human body detection and geolocalization for uav search and rescue missions using color and thermal imagery. In *Proceedings of the Aerospace Conference*, pages 1–8, March 2008.

[52] G. Rufino and A. Moccia. Integrated vis-nir hyperspectral/thermal-ir electro-optical payload system for a mini-uav. *American Institute of Aeronautics and Astronautics 5th Aviation, Technology, Integration, and Operations Conference*, pages 26–29, September 2005.

[53] G. Schaefer, R. Tait, K. Howell, A. Hopgood, P. Woo, and J. Harper. *User Centered Design for Medical Visualization*, chapter Automated overlay of infrared and visual medical images, pages 174–183. IGI Global, 2008.

[54] U. Schilcher, M. Gyarmati, C. Bettstetter, Y. W. Chung, and Y. H. Kim. Measuring inhomogeneity in spatial distributions. In *Proceedings of Vehicular Technology Conference (VTC Spring)*, pages 2690–2694, May 2008.

[55] H. Schultz, A. R. Hanson, E. M. Riseman, F. Stolle, Z. Zhu, C. D. Hayward, and D. Slaymaker. A system for real-time generation of geo-referenced terrain models. In *Proceedings of SPIE Symposium on Enabling Technolgies for Law Enforcement*, 2000.

[56] A. Sehgal, D. Cernea, and M. Makaveeva. Real-time scale invariant 3d range point cloud registration. In *Proceedings of the 7th international conference on Image Analysis and Recognition - Volume Part I*, pages 220–229, Berlin, Heidelberg, 2010. Springer-Verlag.

[57] H.-Y. Shum and R. Szeliski. Construction and refinement of panoramic mosaics with global and local alignment. In *Proceedings of Sixth International Conference on Computer Vision*, pages 953–956, 1998.

[58] G. Sibley, C. Mei, I. Reid, and P. Newman. Adaptive relative bundle adjustment. In *Proceedings of Robotics: Science and Systems (RSS)*, Seattle, USA, June 2009.

[59] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Transactions on Graphics (TOG)*, 25(3):835–846, July 2006.

[60] Z. Song and X. Cheng. A new search engine filtering scheme based on improved neural network and ontology. In *Proceedings of International Conference on Computational and Information Sciences*, pages 178–181. IEEE Computer Society, December 2010.

[61] D. Steedly, C. Pal, and R. Szeliski. Efficiently registering video into panoramic mosaics. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1300–1307, Los Alamitos, CA, USA, October 2005.

[62] S. Suzuki and K. Abe. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985.

[63] R. Szeliski. Image alignment and stitching: a tutorial. *Foundations and Trends in Computer Graphics and Vision Series*, 2(1):1–104, 2006.

[64] R. Szeliski. *Computer Vision: Algorithms and Applications.* Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[65] M. Teke and A. Temizel. Multi-spectral satellite image registration using scale-restricted SURF. In *Proceedings of 20th International Conference on Pattern Recognition (ICPR)*, pages 2310–2313, August 2010.

[66] D. Tingdahl and L. Van Gool. A public system for image based 3d model generation. In *Proceedings of the 5th international conference on Computer vision/computer graphics collaboration techniques*, MIRAGE'11, pages 262–273, Berlin, Heidelberg, 2011. Springer-Verlag.

[67] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, ICCV '99, pages 298–372, London, UK, UK, 2000. Springer-Verlag.

[68] H. J. Trussell and M. J. Vrhel. *Fundamentals of Digital Imaging.* Cambridge University Press, 2008.

[69] V. Vaidehi, R. Ramya, M. Prasannadevi, N. T. Naresh Babu, P. Balamurali, and M. Girish Chandra. Fusion of multi-scale visible and thermal images using EMD for improved face recognition. In *Proceedings of International MultiConference of Engineers and Computer Scientists*, volume I, pages 543–548, 2011.

[70] M. Vollmer and K. Möllmann. *Infrared Thermal Imaging: Fundamentals, Research and Applications.* John Wiley & Sons, 2011.

[71] S. Šegvic. A multimodal image registration technique for structured polygonal scenes. In *Proceedings of Image and Signal Processing and Analysis (ISPA)*, pages 500– 505, 2005.

[72] P. D. Wellner. Adaptive thresholding for the digitaldesk. Technical report, 1993.

[73] D. Wischounig-Strucl, M. Quartisch, and B. Rinner. Prioritized data transmission in airborne camera networks for wide area surveillance and image mosaicking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 17 –24, June 2011.

[74] P. R. Wolf. *Elements of photogrammetry, with air photo interpretation and remote sensing.* International student edition. McGraw-Hill, 1983.

[75] H. Xiang and L. Tian. Method for automatic georeferencing aerial remote sensing (RS) images from an unmanned aerial vehicle (UAV) platform. *Biosystems Engineering*, 108(2):104–113, 2011.

[76] S. Yahyanejad, J. Misiorny, and B. Rinner. Lens distortion correction for thermal cameras to improve aerial imaging with small-scale UAVs. In *Proceedings of IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, pages 231–236, September 2011.

[77] S. Yahyanejad, M. Quaritsch, and B. Rinner. Incremental, orthorectified and loop-independent mosaicking of aerial images taken by micro UAVs. In *Proceedings of IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, pages 137–142, September 2011.

[78] S. Yahyanejad and J. Strom. Removing motion blur from barcode images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 41–46, Los Alamitos, CA, USA, 2010.

[79] S. Yahyanejad, D. Wischounig-Strucl, M. Quaritsch, and B. Rinner. Incremental mosaicking of images from autonomous, small-scale UAVs. In *Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 329–336, 2010.

[80] C. Yuanhang, H. Xiaowei, and X. Dingyu. A mosaic approach for remote sensing images based on wavelet transform. In *Proceedings of the Fourth International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM)*, pages 1–4, 2008.

[81] W. Yue, W. Yun-dong, and W. Hui. Free image registration and mosaicing based on tin and improved szeliski algorithm. In *Proceedings of International Society for Photogrammetry and Remote Sensing (ISPRS)*, volume XXXVII, Beijing, 2008.

[82] Y. Zhan-long and G. Bao-long. Image registration using rotation normalized feature points. In *Proceedings of the Eighth International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 237–241, Washington, DC, USA, 2008. IEEE Computer Society.

[83] G. Zhou. Geo-referencing of video flow from small low-cost civilian uav. *IEEE Transactions on Automation Science and Engineering*, 7(1):156–166, 2010.

[84] Z. Zhu, E. M. Riseman, A. R. Hanson, and H. J. Schultz. An efficient method for geo-referenced video mosaicing for environmental monitoring. *Machine Vision and Applications*, 16(4):203–216, 2005.

Appendix

List of Abbreviations

DEM digital elevation model, p. 21

FIR far infrared, p. 7

FOV field of view, pp. 7, 12, 43, 44

GCP ground control point, pp. 30, 32

GPS global positioning system, pp. 2–4, 8, 15, 19, 20, 22, 25, 26, 29, 30, 57

IMU inertial measurement unit, pp. 3, 4, 15, 19, 20, 22, 25, 29

IR infrared, pp. 2, 6, 7, 38–41

MIR mid infrared, p. 7

MSE mean squared error, p. 68

MWIR mid-wave infrared, pp. 23, 43

NEdT noise equivalent differential temperature, p. 43

NIR near infrared, pp. 7, 22, 23

POV point of view, p. 12

QoS quality of service, p. 18

RANSAC RANdom SAmple Consensus, pp. 11, 33, 47

RFAE robust features along the edge, pp. 49, 55, 68, 70–72, 78

SIFT scale-invariant feature transform, pp. 11, 31, 45–48, 70, 71, 78

SURF speeded up robust features, pp. 11, 19, 21, 31, 45, 47–49, 51, 53, 64, 70, 71, 75, 78

UAV unmanned aerial vehicle, pp. 1–5, 8, 12, 15–20, 22, 24–26, 29–32, 35, 38, 44, 47, 49–51, 55, 64, 78, 79