Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 871

Xin Wang

Robust and Fast Global Image Orientation

München 2021

Verlag der Bayerischen Akademie der Wissenschaften

ISSN 0065-5325 | ISBN 978-3-7696-5283-3

Diese Arbeit ist gleichzeitig veröffentlicht in:

Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover

ISSN 0174-1454, Nr. 373, Hannover 2021

Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 871

Robust and Fast Global Image Orientation

Von der Fakultät für Bauingenieurwesen und Geodäsie
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades
Doktor-Ingenieur (Dr.-Ing.)
genehmigte Dissertation

Vorgelegt von

M. Sc. Xin Wang

Geboren am 12.18.1989 in Hubei, China

München 2021

Verlag der Bayerischen Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5283-3

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover
ISSN 0174-1454, Nr. 373, Hannover 2021

Adresse der DGK:

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften (DGK)

Alfons-Goppel-Straße 11 ● D – 80 539 München

Telefon +49 – 331 – 288 1685 ● Telefax +49 – 331 – 288 1759

E-Mail post@dgk.badw.de ● http://www.dgk.badw.de

Prüfungskommission:

Vorsitzender: Prof. Dr. Philipp Otto

Referent: Prof. Dr.-Ing. habil. Christian Heipke

Korreferenten: Prof. Dr.-Ing. Helmut Mayer (Universität der Bundeswehr München)
Prof. Dr.-Ing. Steffen Schön

Tag der mündlichen Prüfung: 18.05.2021

ISSN 0065-5325 ISBN 978-3-7696-5283-3

Abstract

The estimation of image orientation (also called pose) has always played a crucial role in the field of photogrammetry since it is a fundamental prerequisite for the subsequent works of multi-view dense matching, generating DEM and DSM, etc. In the community of computer vision, the task is also well known as Structure-from-Motion (SfM), which reveals that image pose, while positions of object points are determined interdependently. Despite a lot of efforts over the last decades, it has recently gained the photogrammetrists' interests again due to the fast-growing number of different resources of images. New challenges are posed for accurately and efficiently orienting various image datasets (e.g., unordered datasets with a large number of images, or images compromised of critical stereo pairs).

In this thesis, the relevant ambition is to develop a new fast and robust method for the estimation of image orientation which is capable of coping with different types of datasets. To achieve this goal, the two most time-consuming steps of image orientation are in particular taken care of: (a) image matching and (b) the estimation process. To accelerate the image matching process, a new method employing a random k-d forest is proposed to quickly obtain pairs of overlapping images from an unordered image set. After that, image matching and the estimation of relative orientation parameters are performed only for pairs found to be very likely overlapping. On the other hand, to estimate the image poses in a time efficient manner, a global image orientation strategy is advocated. Its basic idea is to first simultaneously solve all available images' poses, before a final bundle adjustment is carried out once for refinement. The conventional two-step global approach is pursued in this work, separating the determination of rotation matrices and translation parameters; the former is solved by an existing popular method of Chatterjee and Govindu [2013], and the latter are estimated globally using a newly developed method: translation estimation integrating both the relative translations and tie points. Tie points within triplets are adopted to firstly calculate global unified scale factors for each available pairwise relative translation. Then, analogous to rotation estimation, translations are determined by performing an averaging operation on the scaled relative translations.

In order to improve the robustness of the solution, efforts in this thesis are also focused on coping with outliers in the relative orientations (ROs), which global image orientation approaches are particularly sensitive to. A general method based on triplet compatibility with respect to loop closure errors of relative rotations and translations is presented for detecting blunders in relative orientations. Although this procedure eliminated many gross errors in the input ROs, it typically cannot sort out blunders which are caused by repetitive structures and

critical configurations, such as inappropriate baselines (very short baseline or baselines parallel to the viewing direction). Therefore, another new method is proposed to eliminate wrong ROs which have resulted from repetitive structures and very short baselines. Two corresponding criteria that indicate the quality of ROs are introduced. Repetitive structure is detected based on counts of conjugate points of the various image pairs, while very short baselines are found by inspecting the intersection angles of corresponding image rays. By analyzing these two criteria, incorrect ROs are detected and eliminated. As correct ROs of image pairs with a wider baseline nearly parallel to both viewing directions can be valuable, a method to identify and keep these ROs is also a part of this research.

The validation and evaluation of the proposed method are thoroughly conducted on various benchmarks including ordered and unordered sets of images, images with repetitive structures and inappropriate baselines, etc. In particular, robustness is investigated by demonstrating the efficacy of the corresponding RO outlier detection methods. The performance and time efficiency of determining image orientation are evaluated and compared with several state-of-the-art global image orientation approaches.

In summary, based on the experimental results, the developed methods demonstrate to be able to accomplish the image orientation task fast and robustly on different kinds of datasets.

Keywords global image orientation, global rotation estimation, global translation estimation, image matching, relative orientation outliers, repetitive structures, very short baselines

Zusammenfassung

Die Berechnung der Bildorientierung (auch als Bildposenschätzung bezeichnet) hat auf dem Gebiet der Photogrammetrie immer eine entscheidende Rolle gespielt, da sie eine Grundvoraussetzung für darauf aufbauende Arbeiten, wie die dichte Bildzuordnung oder die Generierung von DEMs und DSMs, bildet. Insbesondere im Bereich der Computer Vision ist die Bildorientierung auch unter dem Begriff Structure-from-Motion (SfM) bekannt, was deutlich macht, dass die Orientierung eines Bildes und die Positionen von Objektpunkten voneinander abhängig bestimmt werden. Obwohl bereits seit Jahrzehnten unterschiedlichste Anstrengungen in diesem Themengebiet unternommen werden, hat es in letzter Zeit, aufgrund der schnell wachsenden Anzahl unterschiedlicher Bildressourcen, wieder das Interesse der Photogrammeter geweckt. Es ergeben sich neue Herausforderungen für die genaue und effiziente Orientierung von Bildern aus verschieden Datensätzen (z.B. ungeordnete Datensätze mit einer großen Anzahl an Bildern oder Bilder, die durch kritische Stereopaare beeinträchtigt werden).

Im Rahmen dieser Arbeit wird hierzu eine neue schnelle und robuste Methode zur Bildorientierung entwickelt, die mit verschiedenen Arten von Datensätzen umgehen kann. Um dieses Ziel zu erreichen, werden insbesondere die beiden zeitaufwändigsten Arbeitsschritte der Bildorientierung genauer betrachtet: (a) Bildzuordnung und (b) Bildposenschätzung. Zur Beschleunigung der Bildzuordnung wird ein neues Verfahren unter Verwendung eines so genannten random k-d forest vorgeschlagen, mit welchem Paare aus überlappenden Bildern schnell aus einer ungeordneten Menge an Bildern extrahiert werden können. Anschließend werden die Bildzuordnung und die Schätzung der relativen Orientierungsparameter nur für Paare durchgeführt, bei denen eine Überlappung sehr wahrscheinlich ist. Zur zeiteffizienten Schätzung der Bildposen wird ein globaler Bildorientierungsansatz verwendet. Die Grundidee besteht darin, zunächst alle verfügbaren Bildposen initial global zu bestimmen, bevor zum Abschluss einmalig eine Bündelausgleichung zur Verfeinerung der initialen Posen durchgeführt wird. In dieser Arbeit wird der herkömmliche zweistufige globale Ansatz genutzt, bei dem die Bestimmung von Rotationen und Translationsparametern getrennt voneinander betrachtet werden. Erstere werden durch eine existierende, häufig verwendete Methode von Chatterjee und Govindu [2013] bestimmt, während letztere global geschätzt werden. Dazu werden die Translationsparameter mit Hilfe von relativen Translationen und Verknüpfungspunkten bestimmt. Verknüpfungspunkte in Form von Tripletts werden verwendet, um zunächst global einheitliche Skalierungsfaktoren für jede verfügbare paarweise

relative Translation zu berechnen. Anschließend werden, analog zur Rotationsschätzung, die Translationen bestimmt, indem das Mittel der skalierten relativen Translationen bestimmt wird.

Um die Robustheit der Lösung zu erhöhen, befasst sich ein Teil dieser Arbeit mit dem Umgang von Ausreißern in den relativen Orientierungen (ROs), gegenüber welchen globale Bildorientierungsansätze besonders empfindlich sind. Es wird eine allgemeine Methode zur Erkennung von Fehlern in relativen Orientierungen vorgestellt, die auf der Triplett-Kompatibilität (in Bezug auf Schleifenschlussfehler von relativen Rotationen und Translationen) basiert. Obwohl diese Methode in der Lage ist, die meisten der gegebenen ROs zu bereinigen, kann sie typischerweise keine Fehler erkennen, die auf sich wiederholende Strukturen und kritische Konfigurationen zurückzuführen sind, wie z.B. ungeeignete Basen (sehr kurze Basen oder Basen parallel zur Blickrichtung). Daher wird eine weitere neue Methode vorgeschlagen, um falsche ROs zu eliminieren, die durch sich wiederholende Strukturen und sehr kurze Basen entstehen. Es werden hierzu zwei Kriterien eingeführt, die die Qualität der ROs bewerten. Sich wiederholende Strukturen werden anhand der Anzahl an korrespondierenden Punkten der verschiedenen Bildpaare erkannt, während sehr kurze Basen durch die Schnittwinkel der entsprechenden Bildstrahlen identifiziert werden. Durch die Analyse dieser beiden Kriterien werden falsche ROs erkannt und eliminiert. Da korrekte ROs von Bildpaaren mit einer längeren Basis, die nahezu parallel zu beiden Blickrichtungen verläuft, vorteilhaft für die weitere Berechnung sein können, ist eine Methode zur Identifizierung und Beibehaltung solcher ROs ebenfalls Teil dieser Forschung.

Die Validierung und Evaluierung der vorgeschlagenen Methode wird auf verschiedenen Benchmarks durchgeführt, darunter geordnete und ungeordnete Bildsätze, Bilder mit sich wiederholenden Strukturen und ungeeigneten Basen. Insbesondere die Robustheit der vorgestellten Methodik wird untersucht, indem die Wirksamkeit der Erkennung von ROs-Ausreißern demonstriert wird. Abschließend werden Leistungsfähigkeit und Zeiteffizienz bei der Bestimmung der Bildorientierung evaluiert und mit mehreren State-of-the-Art-Ansätzen zur globalen Bildorientierung verglichen.

Zusammenfassend zeigen die experimentellen Ergebnisse, dass die entwickelten Methoden in der Lage sind, die Aufgabe der Bildorientierung auf unterschiedlichen Arten von Datensätzen schnell und robust zu bewältigen.

Schlagworte globaler Bildorientierungsansatz, globale Rotationsschätzung, globale Translationsschätzung, Bildzuordnung, Ausreißern in den relativen Orientierungen, wiederholende Strukturen, sehr kurze Basen

Acronyms and Table of Notations

General Acronyms

ANN	Approximate Nearest Neighbor
ADMM	Alternating Direction Method of Multipliers
DEM	Digital Elevation Model
DLT	Direct Linear Transformation
DoG	Difference-of-Gaussian
DSM	Digital Surface Model
EG	Epipolar Geometry
GNSS	Global Navigation Satellite System
GPS	Global Position System
GPU	Graphics Processing Unit
IMU	Inertial Measurement Unit
MST(s)	Minimum Spanning Tree(s)
PB	Photogrammetric Block
RANSAC	Random Sample Consensus
ROs	Relative Orientations
SfM	Structure from Motion
SLAM	Simultaneous Localization and Mapping
SDP	Semidefinite Program
$SO(3)$	Lie group denoted as Special Orthogonal Group (3)
$\mathfrak{so}(3)$	Li algebra of $SO(3)$
SVD	Singular Value Decomposition
VO	Visual Odometry

General Notations

$\boldsymbol{C_i}$	Projection Center of i-th image in a global unified coordinate system
$\boldsymbol{I_{3\times 3}}$	Identity Matrix of size of 3×3
N	Number of input images
N_e	largest number of connected images in a photogrammetric block
$\boldsymbol{R_{ij}}$	Matrix of relative rotation between i-th image and j-th image
$\boldsymbol{R_i}$	Rotation Matrix of i-th image in a global unified coordinate system
$\boldsymbol{t_{ij}}$	Vector of relative translation between i-th image and j-th image
$\boldsymbol{x_i}$	Homogeneous image coordinates

Preprocessing

Acronyms

$\boldsymbol{AM}$	Adjacency Matrix
BPVD	Baseline nearly Parallel to Viewing Directions
CDF	Cumulative Distribution Function
$\boldsymbol{E}$	Essential Matrix
$\boldsymbol{FP}$	Matching Feature point set from all images
$\boldsymbol{F}$	Fundamental Matrix
$\boldsymbol{H}$	Homography Matrix
$\boldsymbol{PN}$	Parent Node of a k-d tree
$\boldsymbol{PN_i^j}$	The *j*-th node from the *i*-th layer of a k-d tree
Qxx	Cofactor of an object point's X coordinate
Qyy	Cofactor of an object point's Y coordinate
Qzz	Cofactor of an object point's Z coordinate
RS	Repetitive Structures
VSB	Very Short Baseline

Notations

a	Free parameter for determining potential overlapping pairs
$avg\theta$	Measurement for distinguishing VSB and BPVD pairs
b	Minimum percentage of conjugate points after EG verification
BL_{ij}	Indicator for the degree of VSB or BPVD of image pair (i, j)
cp_{min}	Threshold for the minimum number of conjugate points for an overlapping pair
d_{ij}	Scalar Product of *i*-th target feature and *j*-th retrieved feature
d	Threshold for rejecting d_{ij}
D_{ij}	Average of all available values of d_{ij}
N_c	Minimum Number of conjugate points after EG verification
n_tr	Number of random k-d trees
P_{ij}	Number of retrieved nearest features for image pair (i, j)
RS_{ij}	Measurement for the degree of RS of the *i*-th and *j*-th images
nRS_{ij}	Normalized Measurement of RS_{ij}
S_{ij}	Measurement for Image Similarity

Global image orientation methods

Acronyms

IRLS	Iteratively Reweighted Least Squares
L1RA	Lie-Algebraic global rotation estimation using L_1 norm
RBA	Robust bundle adjustment

Notations

$\mathcal{G}$	Viewgraph generated after preprocessing
$\mathcal{V}$	Set of vertices in $\mathcal{G}$
$\mathcal{E}$	Set of edges in $\mathcal{G}$
$\boldsymbol{R}$	3D rotation matrix
$\boldsymbol{\varrho}$	Lie algebra vectorized representation of Lie group element $SO(3)$ $\boldsymbol{R}$
$\boldsymbol{A_v}$	Coefficient matrix of solving global rotation
ϵ	Terminal criterion of solving global rotation
$\rho(x)$	Huber-like loss function of solving global rotation
$\phi(x)$	Function for iteratively reweighting global rotation
X_0, Y_0, Z_0	3 dimensional translation parameters
X, Y, Z	3 dimensional coordinates of object points
ε_r	Maximum allowed value of triplet rotational compatibility
ε_t	Maximum allowed value of triplet translational compatibility
λ_{ij}	Global unified consistent scale factor for image pair (i, j)
η_{ij}^{i}	Consistent scale factor of each image pair (i, j) in the reference of image i
r_{jk}^{i}	Consistent scale factor which is constant with respect to the corresponding triplet of images (i, j, k)
$\boldsymbol{A}_Z$	Coefficient matrix for solving all tuples with respect to a specific image
γ_i	Scale factor for transfer tuples into global unified system
$\boldsymbol{A}_R$	Coefficient matrix for transferring all tuples into global unified system
$\boldsymbol{A}_P$	Coefficient matrix for global translation estimation using relative translations
$\boldsymbol{X}_P$	Unknown global translation parameters
$\boldsymbol{K_i}$	Intrinsic calibration matrix of image i
$\varphi(x)$	Back projection function of collinearity equation
$f_h(x)$	Huber loss function of RBA
ϵ_{ba}	Terminal criterion of RBA
T_{rba}	Maximum number of iterations in RBA
v_r	Threshold of residual for eliminating blunder observations in RBA
d_a	Threshold of minimum intersection angle for eliminating tie points in RBA
T_{op}	Threshold of minimum number of tie points visible in an image in RBA

Contents

1 Introduction

Over the centuries, 2D art have always played a particularly important role in human civilization. Artists like *Raffaello Sanzio da Urbin (1483-1520)* and *Zhang Zeduan (1085-1145)* left us many priceless 2D paintings, such as "*The school of Athens*" and "*Along the River During the Qingming Festival*", which typically depict some realistic, but non-existent 3D scenarios. Putting a lot of effort on improving the artistry, geometric knowledge is often only gathered empirically in such earlier works. With the emergence of photography in the 1830s, naturalism in art, which was mainly supported since the sixteenth century by the knowledge of the perspective projection, was substituted by the technique of aligning perspective images as photographs in the nineteenth century. Images were then popularly used as an inspiration to perceive environment, and of course also for documentation, reconnaissance and surveillance. However, paintings and images are both 2D artifacts with uncertain depth information. With the development of methods for 2D image processing, photographs did not only initiate the transition to modernity in art, but were also applied to solve engineering problems, such as the 3D mensuration of buildings for preserving cultural heritage [Albertz, 2001; Remondino et al., 2016].

The determination of 3D information (which additionally contains depth information) from 2D images which observe the same scene from different viewpoints is a fundamental task in photogrammetry. An essential pre-required step of this task is to determine the image poses or image orientations, which describe where the images got exposed and the viewing direction, as well as the 3D coordinates of tie points, and thus a sparse 3D point cloud. This process is often called structure-from-motion (SfM) in computer vision and simultaneous localization and mapping (SLAM) in robotics. Measuring based on images dates back to the invention of photography and was further investigated by scientists from the field of optics (Ignazio Porro and Ernst Abbe), while, later on, stereo photogrammetry was used for terrestrial mapping applications on mountain areas. In the time of analogue and analytical photogrammetry, airborne and terrestrial images were processed with comparators and plotters for map generation.

Following the development of digital sensor and electronic information technology, the application of photogrammetry was extended from professional photogrammetrists to common users, for example, making 3D models of complete cities as shown in Google Maps and allowing to virtually explore tourist sites in 3D when people are not able to go there physically (e.g. during a quarantine - the Große Kuppelhalle Bode Museum provided an online virtual site

during the Covid-19 period)[1]. Nowadays, digital images are omnipresent, as the cost for such an image is pretty low and even a user-level cellphone is capable of generating images of relatively high quality (for instance, one of the most popular cellphone, iPhone 11, has two cameras with 12 million pixels each). In addition, people often share their pictures on some open websites or social media applications, such as Flickr and Facebook, making it easy to access a wide range of pictures. Although these pictures were initially not taken for photogrammetric purposes, it is of great interest to carry out reconstruction or measuring tasks on these abundant pictures. A prominent use case for such tasks is the reconstruction of Notre Dame de Paris. Notre Dame de Paris, which is one of the most famous churches around the world, got damaged by an unexpected fire on April 16th, 2019. In consequence, the French government decided to rebuild it. One possibility to provide a geometric reference for the task of rebuilding is the usage of a point cloud together with the image poses determined based on touristic pictures from the Internet. Figure 1.1 shows such a reconstruction example of Notre Dame de Paris using pictures collected by Snavely et al. [2006] which is processed by the method presented in this thesis.

In addition to the example of 3D scene reconstruction, images are also successfully applied in navigation and automatous driving. One strategy is to embed a camera system as an auxiliary sensor to allow for continuous navigation even when GPS and GNSS (Global Navigation Satellite System) signal deny case occurs, for example, when an automatous vehicle enters a tunnel where GPS and GNSS signals can no longer be received for navigation, the location of the vehicle can be obtained by computing the position of the mounted camera.

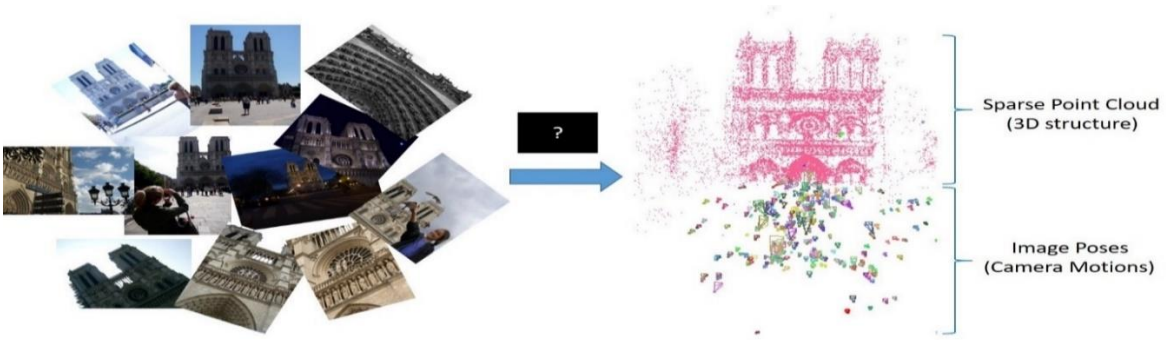


Figure 1.1.: Collection of pictures of Notre Dame de Paris [Snavely et al., 2006; Wilson and Snavely, 2014] and the corresponding reconstruction of the sparse point cloud and image poses using the method proposed in this thesis which is denoted as the black box in this figure.

While images are now much easier to access and capturing datasets is not a main issue anymore, the rapidly increasing computational power and resources resolve the limitation of working only with small datasets. Therefore, an efficient solution for being capable to deal with reconstructions from a large amount of images is in demand. In particular, such a solution should allow to orient a lot of images as well as reconstruct large areas by taking time efficiency, robustness and accuracy into account.

[1] http://bode360.smb.museum/?from=timeline

Modern methods usually solve the image orientation problem via bundle adjustment techniques, which optimize a cost function known as the total reprojection error, where the reprojection error is calculated as the discrepancy of measured image feature point coordinates and the corresponding coordinates estimated using the well-known collinearity equations. The term "bundle adjustment" originates from the fact that optical rays from different viewpoints to the same objects are adjusted by iteratively fine-tuning the positions of object points and the image poses such that the total reprojection error is minimized under a certain statistical model. While the concept was developed more than 60 years ago Schmid [1958]. An excellent survey on bundle adjustment is provided by Triggs et al. [2000]. The complexity of this minimization approach originates from the nonlinearity of the collinearity equations and the non-convexity of the optimization function, making it impossible to obtain an optimal solution for bundle adjustment directly. Thus, initial values are an essential input for solving bundle adjustment and the quality of these initial values is normally of great significance to avoid local minima.

To solve the image orientation task, according to the procedure in which images are oriented, there are mainly three different strategies: incremental, hierarchical and global methods[2]. The ***incremental*** approach [Snavely et al., 2006; Agarwal et al., 2009; Schönberger and Frahm, 2016; Wu, 2013; Wang et al, 2018 and 2019a] is the earliest and most intuitive idea. Two images or triplets are initially chosen according to some specific requirements; their relative orientation parameters are computed; new images are iteratively added to extend the photogrammetric block by space resection (also called *PnP* or perspective-n-point problem) and triangulation; a robust bundle adjustment is typically adopted to obtain refined results. Farenzena et al. [2009], Mayer [2014] and Toldo et al. [2015] present a so-called ***hierarchical*** method, which improves the incremental idea by first dividing the images into overlapping subsets, and then processing all subsets individually by incremental SfM. Finally, the subsets are merged in a hierarchical way with a number of bundle adjustments. Both strategies are relatively slow because of the repeated use of bundle adjustments. To overcome this problem, the ***global*** method [Govindu, 2001; Martinec & Pajdla, 2007; Jiang et al., 2013; Moulon et al., 2013; Ozyesil et al., 2015; Arrigoni et al., 2016; Reich & Heipke 2016; Goldstein et al., 2016; Wang et al., 2019a and 2019b] considers this problem from a different perspective. Global methods build on the well-known idea that rotation and translation estimation (i.e. the computation of the 3D coordinates of the projection center) can be separated. Accordingly, these methods consist of two main steps: global rotation averaging and global translation estimation. Global rotation averaging simultaneously estimates the rotation matrices of all available images in a consistent (global) coordinate system [Hartley et al., 2013]. Given global rotations, global translation estimation aims at simultaneously solving the translation parameters of all available images. The advantage of global SfM is that it can solve both, rotations and translations, without intermediate bundle adjustments, only a final one is necessary. However, this method is more sensitive to outliers than the two others.

[2] Global pose estimation methods consist of global translation and global rotation determination. In some cases, the term "global" is used in the context of optimization. In this thesis, the term "global" is applied to refer to approaches that take into account the information about relative poses of all overlapping image pairs simultaneously.

In this thesis, efforts are concentrated on developing a time efficient and robust image orientation approach, which contains methods for detecting mutual overlapping image pairs, relative orientation (RO) outlier elimination and global image orientation. In the next subsection, the approach is motivated, the objective is formulated and the corresponding characteristics are outlined.

1.1 Motivation and objective

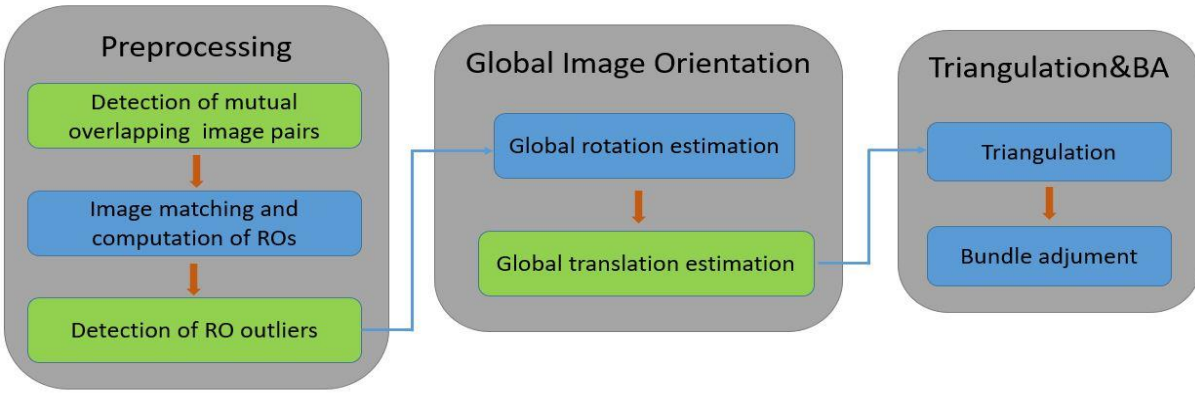


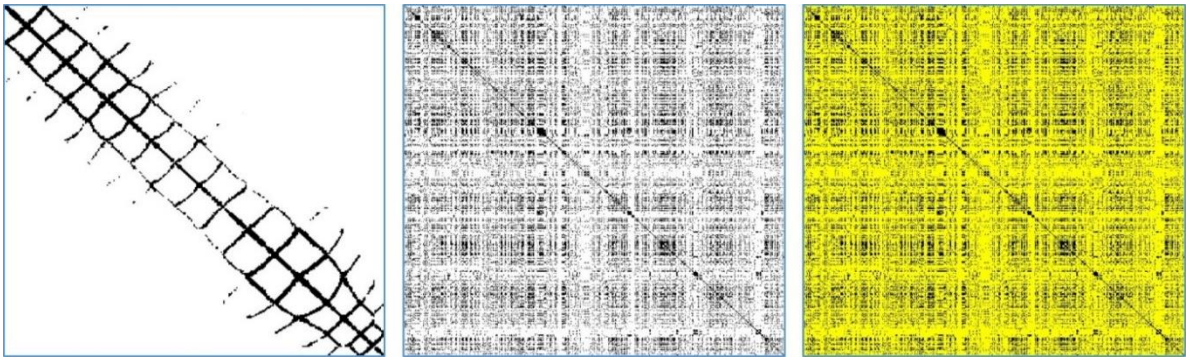
Figure 1.2.: Overview of the workflow presented in this thesis.

Thanks to the development of photogrammetry and multi-view geometry in computer vision, the estimation of image orientation parameters or SfM has been well studied in the last several decades, which can be demonstrated by some successful academic and commercial packages, such as *VisualSFM*, *Colmap* and *Photoscan*[3]. In consequence, one may doubt and ask *is it still necessary to spend further time and effort on this topic*? The answer is a clear Yes - research on image orientation is still an ongoing matter. This positive answer is given according to arguments which are presented in detail in the next paragraphs and are intuitively illustrated in Figure 1.2.

As Figure 1.2 shows, the method proposed in this work is composed of three serial and interdependent stages. In the initial *preprocessing* stage, mainly two tasks are addressed: mutual overlapping image pairs are detected first to improve the time efficiency of image matching and the computation of relative orientation. Secondly, outliers of relative orientations are handled to improve the robustness. Based on these inliers pairs of relative orientation, the second stage is denoted as *global image orientation* which aims to estimate the exterior orientation parameters of all available images simultaneously. For this purpose, the steps of global rotation estimation and global translation estimation are implemented separately. The last stage is *triangulation and bundle adjustment*, in which the coordinates of tie points are obtained by triangulation and the results, including the image orientation parameters and coordinates of tie points, are refined by a robust bundle adjustment. Note that the triangulation & refinement stage

[3] More information about *VisualSFM*, *Colmap* and *Photoscan* can be found by the links of http://ccwu.me/vsfm, https://colmap.github.io/ and https://www.agisoft.com.

is not part of the novelty of this thesis. However, it is an essential part of the whole work, because the calculated initial image pose parameters should not only guarantee the success of triangulation, but should also allow the bundle adjustment to converge. Any violation of these two requirements in turn indicates the failure of the solution of initial value computation. In addition, the refined accurate image poses and coordinates of object points are often used in subsequent multi-view stereo processing and DSM (Digital Surface Model) or DEM (Digital Elevation Model) generation.



(a) UAV images (b) Unordered images (c) Exhaustive matching

Figure 1.3.: Overlap graph of ordered and unordered datasets, respectively. The vertical and horizontal axes are the image IDs from 1 to N, these overlap graphs are typically symmetric. Black pixels mean that the corresponding two images overlap, while white ones denote a non-overlapping relationship between image pairs. Yellow pixels are extra image pairs (beyond the determined black ones) that need to be matched using an exhaustive image matching strategy.

As feature extraction and feature matching are the standard preliminary works for almost every feature-based image orientation method, which is also true for this thesis, one bottleneck of such preliminary works is the effort needed, growing quadratically with the number of images to be solved. A conventional way is to carry out $N(N\text{-}1)/2$ exhaustive pairwise image matchings [Snavely et al., 2006; Olsson and Enqvist, 2011], where N is the number of images. In aerial photogrammetry or a pre-planned image capture project, things become much easier, because overlapping image pairs can be determined in advance, either by the position information from GPS/IMU (Inertial Measurement unit) or by prior knowledge on the pre-planned pattern. In contrast, in close-range photogrammetry, untrained users typically do not capture images in a pre-planned pattern, and even experts often need significantly more time for data acquisition, when strict recording protocols must be followed. Moreover, as it has already been addressed, images taken without a pre-planned pattern can be obtained in a much cheaper way, e.g., when considering crowd-sourced data from the Internet. However, datasets acquired in such a way are usually unordered[4] and the quadratic image matching effort must be taken into account. The differences between these various acquisition approaches are exemplarily shown in Figure 1.3: The overlap graph of the ordered dataset from a UAV platform in (a) shows regularly distributed dark pixels which reflect some basic rules of the flying routes, such as, there are 18 strips in which the last three strips are shorter than the others and every strip overlaps with at least two other strips. The black pixels in the overlap graph of

[4] In this thesis, “Unordered” means that images are unorganized and the corresponding overlapping information is totally unknown, while “Ordered” indicates that the overlapping information is already known via some manners.

the unordered dataset (the images of Notre Dame shown in Figure 1.1 are investigated here) shown in (b) are just irregularly distributed, making it challenging to clarify which image pairs overlap if no prior knowledge is given. Ideally, the most efficient approach is to only run image matching on image pairs labelled as black pixels. One typical way to find these black pixels is to use exhaustive image matching. As Figure 1.3 (c) depicts, in this case a lot of additional effort is required on these non-overlapping image pairs denoted as yellow pixels. To free the experts from the bound of protocols and to allow common users to provide image datasets, an efficient method for detecting mutual overlap of large sets of unordered images is needed and thus presented in this work. Specifically, a random k-d forest is first built from the features extracted from all images and overlapping image pairs are detected by fast nearest neighbor search in this forest.

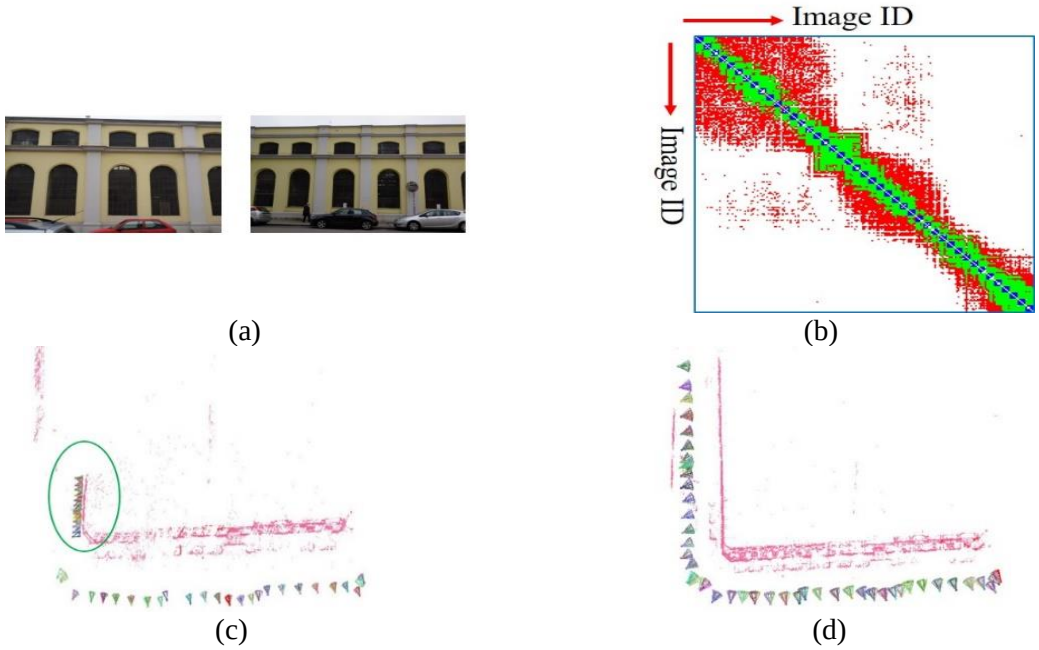


Figure 1.4: An example scene with repetitive structure and image pairs with very short baselines. (a) Two example images with repetitive structure. (b) Ground truth of overlap graph with the image IDs on the horizontal and the vertical axes; green pixels denote overlapping image pairs, red pixels represent non-overlapping pairs with incorrect ROs due to RS, and blue pixels indicate the corresponding VSB image pairs. (c) incorrect reconstruction without eliminating incorrect ROs. (d) accurate reconstruction after eliminating incorrect ROs using the method suggested in this thesis.

Based on the determined overlap graph, the standard image matching process is conducted for conjugate points followed by epipolar geometric validation, in which the RANSAC technique [Fischler and Bolles, 1981] is applied to estimate the essential (or fundamental) matrix. The resulting relative orientation parameters[5] are only considered to be correct if a minimum number of point pairs agree with the model of central perspective. Although some blunders in

[5] In computer vision literature, the term "Epipolar geometry" is often used to describe the relative geometric relationship of one image pair. In this thesis, the term "relative orientation" is instead used. In photogrammetry the result of relative orientation includes relative rotation and translation. These parameters can be derived from the essential (or fundamental) matrix [Longuet-Higgins, 1981; Hartley and Zisserman, 2003] which can be determined from the epipolar geometric constraint.

the estimated ROs can be avoided using this RANSAC filtering step, there are still some incorrect ROs remaining undetected which are normally directly fed into subsequent global image orientation and can therefore stem the proposed global image orientation approach. In this thesis, care is taken of blunders in the ROs that result from problematic observations, repetitive structures and critical configurations such as inappropriate baselines. To obtain a reliable and robust solution, first all possible triplets are extracted where the three images are mutually connected to each other. Incorrect relative orientations are detected by checking the triplet loop closure constraint using both, relative rotation and relative translation. Repetitive structure is a characteristic of a single image and describes the fact that multiple regions of the image look similar. Typically, this is caused by a repetitive 3D structure in the scene (also explaining the naming repetitive structure instead of repetitive texture, as texture refers to the 2D image space). As a consequence, the descriptors of extracted features are rather similar in this case. Matching images with repetitive structure leads to many ambiguous point pairs and outliers. In this context, an *image pair due to repetitive structure* (RS) is referred as non-overlapping RS image pair, for which incorrect conjugate point pairs were extracted due to these ambiguities. Such non-overlapping RS image pairs with nevertheless similarly looking images can stem from, for example, a set of façade images, when the façade is somewhat symmetric. If too many such incorrect point pairs are extracted, it is possible that RANSAC is not be able to detect the error anymore and incorrect relative orientation parameters are derived. A critical configuration with a very short baseline (VSB) results from improper image acquisition planning, e.g. when images are taken in different directions, but from basically the same projection center. In addition, crowd sourced datasets, such as images available on the Internet, are widely used nowadays. These datasets may contain pairs with such critical configurations as well. In Figure 1.4, an example with both, RS and VSB image pairs, is shown. (a) shows repetitive structure of windows and walls, while these two photos are actually showing different walls. (b) depicts the ground truth of the overlap graph where green pixels denote the correct overlapping image pairs; (c) and (d) show the reconstruction result when applying the proposed global SfM method without and with employing the proposed method to eliminate outliers in ROs, respectively. It is obvious that the reconstruction is more reasonable after deleting the incorrect ROs.

1.2 Problem statement and contributions

To improve both, time efficiency and robustness of image orientation, the global strategy is applied in this thesis, as many works have been conducted already to demonstrate the capability of global strategies. In the context of this work, the *global image orientation* task starts with the remaining inliers of relative orientations. Similar to other global orientation approaches, two separate steps of global rotation estimation and global translation estimation are presented in this work.

Global rotation estimation addresses the problem of assigning a rotation matrix to every image in the photogrammetric block in a way that is most consistent with the calculated relative rotations from each remaining RO. In other words, relative rotations $\boldsymbol{R_{ij}}$ are assigned to a set of images, while the global rotations of all available images are computed simultaneously, optimizing for the constraint $\boldsymbol{R_{ij}R_i} = \boldsymbol{R_j}$, where $\boldsymbol{R_i}$ and $\boldsymbol{R_j}$ are the global rotations of *i*-th and *j*-th image. As ample research has been published on global rotation estimation and the problem can be considered to be largely solved, in this thesis an existing method (the one suggested by Chatterjee and Govindu [2013]) is used as a basis for the subsequent novel translation determination. This particular work is chosen for two reasons: first, their work is widely used in many state-of-the-art global image orientation works and is considered capable of providing reliable results for large numbers of images; second, in the experimental evaluation it is important to make a fair comparison of translation results from different methods, which requires to use the same rotation estimation method.

Analogously to rotation estimation, global translation estimation is commonly formulated as computing global image translation parameters from the relative translations of all available image pairs. However, global translation estimation cannot be solved directly from relative translations, because relative translations only contain the normalized translation direction vector between the projection centers of two overlapping images, without providing any information on the length of these vectors. Thus, the problem of estimating global translations is not fully solved in general (in contrast to global rotation averaging), and this thesis develops a novel method for global translation estimation.

In summary, to achieve the objective of solving the image orientation problem for various datasets (including datasets containing ordered and unordered images, or images with repetitive structures and critical configurations) in a time efficient and robust manner, this thesis contains the following main contributions (highlighted in green boxes in Figure 1.2):

- A method to quickly identify mutual overlapping image pairs in large set of unordered images is developed. Without requiring any prior knowledge, time efficiency is increased by carrying out image matching and epipolar geometry computation on these determined overlapping image pairs only.
- An approach to detect blunders in the relative orientations is proposed to improve the robustness of image orientation. RO outliers that are due to noisy observations, repetitive structures and inappropriate baselines are eliminated.
- After deriving global rotation information with the approach of Chatterjee and Govindu [2013], this thesis contributes a new global translation estimation method, using information from tie points and relative translations.
- The capability of the whole pipeline is demonstrated on various datasets (such as, UAV and terrestrial images, images from the Internet, benchmarks from the photogrammetry and computer vision community). All the above mentioned contributions is evaluated and compared to several state-of-art methods.

Parts of this thesis have been published in journal articles as well as in peer-reviewed conference and workshop proceedings. The content of the following chapters is partly adapted

and improved from the author's previous works: [Wang et al., 2017], [Wang et al, 2019b], [Wang et al, 2019c], [Wang and Heipke, 2020].

1.3 Reader's guide

The content of this thesis is structured as follows. Chapter 2 provides a comprehensive overview of related works. A brief review of the endeavors on increasing the efficiency of image matching and ROs outlier detection is followed by a study of existing methods for incremental, hierarchical and global image orientation. In Chapter 3, the preprocessing steps are introduced. The method for the fast identification of mutually overlapping images in large sets of unordered images is explained in detail and the strategy for detecting outliers in ROs is then justified. Chapter 4 addresses the proposed global image orientation method. First, the global rotation estimation method by Chatterjee and Govindu [2013] is briefly presented for the sakes of completeness. Then, a novel global translation estimation method is described in detail. The setup and the datasets used for the experiments are discussed in Chapter 5, before an exhaustive evaluation of the proposed methods is reported in Chapter 6. Results with a focus on time efficiency and robustness are discussed. Finally, Chapter 7 draws conclusions and prospects for future work.

2 State of the art

Over the last several decades, the topics relevant to this thesis have been extensively studied by various communities, gaining lots of achievements. Nowadays, the topic of image orientation faces some new challenges because images can be obtained at very low cost, or even for free, if they are just downloaded from the Internet (e.g., from social websites such as Facebook, Flickr or Instagram). This also implies that methods need to be investigated to process more images in a time efficient and robust manner. [Snavely etc., 2006] is one of the earliest works that deal with large sets of unordered images (especially from the Internet). They presented an incremental pipeline for SfM, demonstrating that it can generate accurate reconstructions in practical scenarios, where hundreds or even thousands of photos captured by different tourists are used as input. The results are refined by using bundle adjustment which is not a convex optimization problem due to the special structure of the exterior orientation parameters of the images. Consisting of $SO(3)$ rotation matrices and $\mathbb{R}^3$ translation vectors, the optimization scheme may converge to an undesired local minimum in realistic settings. It is therefore crucial to develop methods that produce results suitable to initialize the bundle adjustment, meaning that the exterior orientation parameters (and interior orientation parameters) as well as the 3D structures need to be initialized as close as possible to the ground truth. Approaches to estimate these initial values can be categorized into three classes: incremental, hierarchical and global methods.

This chapter is dedicated to give an overview of the existing works that are relevant to this thesis. The following parts of this chapter are structured based on the workflow of the image orientation procedure, which starts with a coarse introduction of some widely used image features and relative orientation methods in Section 2.1 (these aspects are not covered in detail, as this thesis does not tend to make contributions to them). Section 2.2 presents an extensive description of some state-of-the-art methods for efficient image matching. This is followed by a review of relevant works on different classes of solutions to determine image orientation methods: Section 2.3 investigates incremental and hierarchical methods; Section 2.4 presents a comprehensive summary of global methods containing studies to cope with outliers of relative orientations, global rotation estimation, global translation estimation and more recent one-step global image orientation; Section 2.5 discusses some alternative image orientation methods. Finally, this chapter closes with a discussion of the current state of the art and open questions.

2.1 Image features and relative orientation

To derive the desired image orientation parameters, suitable observations need to be identified and extracted from the images and some corresponding preprocessing steps (eg., feature extracting and relative orientation) must be carried out. This section briefly reviews some state-of-the-art studies on these preprocessing steps. By the end of the last century, quite a few researchers developed automatic methods for image feature extraction. In detail, Harris and Stephens [1988] proposed to use image patches with significant grey value variations in two orthogonal directions and Förstner [1986] presented a method using corner points or blobs where the image intensities show a large variation between adjacent regions in all directions. One of the most popular feature extraction techniques is the so-called Scale Invariant Feature Transform (SIFT) [Lowe, 2004]. SIFT can be considered as a blob detector in which an image pattern that differs from its immediate neighborhood with respect to intensity, color and texture, more specifically, features are detected by the maxima and minima of the result of DoG (Difference-of-Gaussian) function applied in scale space to a series of smoothed and resampled images, low-contrast candidate points are discarded, dominant orientations are assigned to localized keypoints. The main advantage of SIFT is its invariance (up to some degree) against rotation, scale, illumination and viewpoint changes. Yet another successful feature called Speeded Up Robust Features (SURF) [Bay et al., 2008], largely inspired by SIFT, employs a box filter to approximate the second order Gaussian and image integrals to compute an image convolution. More recently, many works were published on learning image features [Trzcinski et al., 2005; Chen et al., 2016]. In addition, Chen et al. [2020] have recently shown that in the context of image matching, learned image features are more robust than the conventional hand-crafted features when facing large changes in the viewing angle or affine distortions.

After observations are obtained from the images, the following step for deriving the desired image orientation parameters is to compute pairwise or triplet-wise relative orientations. Ample studies have been carried out on this topic, e.g. dealing with the fundamental or essential matrix [Faugeras, 1992; Hartley,1992] for two views or the trifocal tensor for three views [Hartley, 1997; Ressl, 2000]. Typically, linear estimation systems can be built for solving the corresponding matrix or tensor by using image correspondences and relative orientations can be determined from the entities of the matrix or tensor. However, such linear estimation systems are sensitive to outliers of image correspondences (e.g. spurious matches) and to the degeneracy of specific configurations (e.g., all correspondences are collinear). A common approach is to eliminate outliers by integrating a RANSAC scheme, where a minimal number of required observations are randomly selected to perform an epipolar geometry check. This is repeated for a specified number of iterations and the configuration with the largest set of inliers is denoted as the final solution. In the case of the essential matrix, which is the most relevant to this thesis, the interior orientation parameters of the cameras are known in advance, which reduces the risk

of facing a degenerated case. Currently, the most successful pipeline for estimating an essential matrix is the five-point algorithm proposed by Nistér [2004] and extended by Stewenius et al. [2006]. In the five-point algorithm, an algebraic optimization function is minimized, while outliers and inliers are distinguished based on the geometric reprojection error and an iterative RANSAC scheme is applied for seeking the most robust solution.

2.2 Efficient image matching

Finding correspondences in different images for the task of image orientation is a fundamental step in photogrammetric 3D reconstruction, also referred to as image matching. One way to obtain conjugate points in two or more images corresponding to the same 3D object point is the approximate nearest neighbor (ANN) method based on k-d trees or random k-d forest [Arya et l., 1998; Silpa-Anan, Hartley, 2008; Muja and Lowe, 2009, 2012, 2014]. However, image matching is one of the most time-consuming processes, because, given N images, N*(N-1)/2 image pairs must be matched if no prior knowledge on the image orientations is available. A comprehensive review of the current state of the art, including a comparison of a number of methods is contained in [Hartmann et al., 2016]. Classically, photogrammetry has a preference to work with ordered images [Luhmann et al., 2014], where overlapping image pairs are known prior to the pose estimation. Consequently, only the overlapping image pairs need to be matched. This is typically achieved in one of two ways: either through carefully planned recording, or by measuring coarse image positions during image acquisition using external sensors (e.g. GPS and IMU). To broaden the scope of image-based 3D reconstruction, modern projects may work with crowd-sourced or thousands of unordered images. Therefore, image matching could become the bottleneck – even with today's powerful machines. In this review, three main lines of strategies are inspected: first, the reduction of the number of feature points per image, by identifying those that are most suitable for matching, before the actual image matching is conducted; second, the reduction the number of image pairs, by finding out those which are most likely to overlap and match; third, other integrated methods.

2.2.1 Reduction of the number of features per image

To reduce the number of feature points which are extracted from a given image, the most intuitive way is to modify the underlying feature detector, e.g., by adapting corresponding inherent thresholds. In this review, the DoG is exemplary discussed, however, without loss of generality, the same principles are valid for other feature detectors as well. Two simple ways are naturally adopted: (i) Varying the DoG detection threshold used to decide whether an interest point is considered or not. The SIFT implementation of VLFeat [Vedaldi and Fulkerson, 2008], for example, uses a rather generous value which generates nearly thousand feature points per image on a consumer camera. A stricter threshold will in general return fewer feature points, but with a higher contrast. However, these features are not guaranteed to be more suitable for matching and many useful feature points that have salient textures may be discarded. (ii)

Selecting features in a high scale level. [Wu, 2013], for example, employed the so-called "preemptive matching", in which the extracted SIFT features are sorted in decreasing order with respect to the SIFT scale. Only features in high scales are used for matching.

[Hartmann et al., 2014] approached this problem from another perspective by considering it as a classification problem. The extracted features can be trained to predict whether they are suitable for matching or not. Taking the feature descriptor as input, a binary classifier in form of a random decision forest was trained using a large set of features which have been exhaustively matched., More specifically, the positive training set was generated by the feature points that have at least one match and all the other feature points belonged to the negative set.

2.2.2 Reduction of the number of image pairs

The conventional exhaustive image matching strategy conducts matching for every potential pair of images. However, this strategy is typically contrary to the real conditions because many image pairs just do not overlap due to their viewing angles or their positions of exposure. Therefore, the efficiency of image matching can be significantly improved if the mutual overlap relationships are determined in advance. To approximate the set of actually overlapping image pairs, many works on computing image similarity were proposed: [Nistér and Stewenius, 2006] is one of the earliest studies on this topic. The key assumption of this study is that across all images homologous features should appear similar, which is exploited by quantizing the feature descriptors on a specific indexing structure. Tree structures are used for retrieval, while the k-means algorithm is recursively employed to quantize all the features. This process can be carried out hierarchically until a pre-specified level of detail is reached; in this way a so-called vocabulary tree is created. Each cluster of the vocabulary tree is regarded as one word. It is intuitive that matchable points should be classified into the same word and unmatchable points should be located in different words. To measure image similarity, a weighting scheme is introduced. Using this scheme, an image is represented by a histogram of visual words weighted by *tf-idf* (term frequency – inverse document frequency). This kind of weighting ensures that words appearing seldom have a larger weight [Sivic and Zisserman, 2003], and image similarity can be efficiently computed with the Euclidean distance between the corresponding *tf-idf* vectors. This idea is often referred to as "Bag of Word" (BoW) and is widely used in the context of loop closure in SLAM [Mur-Artal et al., 2015]. To improve efficiency and robustness, [Zhan et al., 2015] extended this method by constructing multi-vocabulary trees implemented on graphic processing units (GPUs). For this purpose, different vocabulary trees are built and each word is evaluated using the average distance between every feature and its cluster center. Later, instead of using local image features, deep convolutional features from a pretrained VGG-16 network [Simonyan and Zisserman, 2014] were used by [Wan et al., 2018]. A certain number of deep convolutional features are extracted from every image and the corresponding vocabulary tree and *tf-idf* vectors are generated using the ideas of Nistér and Stewenius [2006] and Sivic and Zisserman [2003]. The similarity of two images is obtained using the cosine of the angle between two corresponding *tf-idf* vectors. More recently, [Zhan et al., 2018] extracted global features for each image from AlexNet-FC7 (fully connected layers) [Krizhevsky et al.,

2012] and ResNet101-Pool5 (pooling layers) [He et al., 2015], while the cosine similarity was computed using these global features. [Michelini and Mayer, 2020] estimated similarities between images using the Jaccard index as a relative score to rank the images. The Jaccard index is computed as the ratio between the number of matching features of two images and the number of unique features extracted from these two images.

Knowing the image similarity scores only, it is still challenging to distinguish between overlapping and non-overlapping image pairs. Most of the above mentioned methods select a fixed number of pairs with highest similarity scores, which may result in either many redundancies or an insufficient number of image pairs. To overcome this limitation, [Jiang and Jiang, 2020] proposed a method which adapts the similarity threshold by analyzing statistical information of the similarity scores. As a consequence, more image pairs are selected by expanding minimum spanning trees (MST) to avoid the photogrammetric block breaking apart.

2.2.3 Other integrated methods

[Havlena and Schindler, 2014] proposed a method named VocMatch which is again inspired by the idea of vocabulary trees. A 2-level vocabulary tree was built in a way that the first level consists of 4096 clusters and the second level of 4096 * 4096 clusters. The basic idea is that all features of all images are indexed in the resulting set of about 16 million words, instead of matching them to each other in a pairwise fashion. It is assumed that features which are clustered into the same word in the second level represent matchable points, which means that if two cluster centers are located closely to each other, the results may be ambiguous. The authors demonstrated that the complexity reduces from quadratic to linear in the number of images. [Schönberger et al., 2015a] presented a pairwise image geometry encoding pipeline which takes into account the distribution of feature location and orientation and uses a randomforest classifier to predict potentially overlapping image pairs. Later, they further improved their pipeline by exploring the quality of relative geometric configurations and by using a similar random forest predictor to classify good and bad relative orientations [Schönberger et al., 2015b]. Instead of using SIFT features, [Frahm et al., 2010] used global appearance gist features [Oliva and Torralba, 2001] and the k-medoids algorithm with Hamming distance to generate clusters of images. Within each cluster, SIFT features are extracted and matched, then pairwise relationships between clusters are established to ensure the block is connected. [Heinly et al., 2015] improved the idea of [Frahm et al., 2010]. In their approach, clusters are represented by sets of visual words and overlapping image pairs are detected as k-nearest neighbors using a vocabulary tree.

2.3 Incremental and hierarchical image orientation

In this section, related works on two important strategies for the task of image orientation are reviewed: incremental and hierarchical image orientation. Incremental image orientation starts with an initial subset of images, e.g., initializing a small reconstruction, and iteratively adds

further images to the block, running repetitive intermediate bundle adjustment to refine the results. Hierarchical image orientation improves the incremental idea by first dividing the images into overlapping subsets, before processing all subsets individually by incremental image orientation. Finally, the subsets are merged in a hierarchical way with a number of bundle adjustments. In particular, the incremental approach can be indicated as common practice in photogrammetry, outlined in the relevant textbooks (e.g. Kraus [1997], pp.48, Hartley and Zisserman [2003], from pp.435).

2.3.1 Incremental image orientation

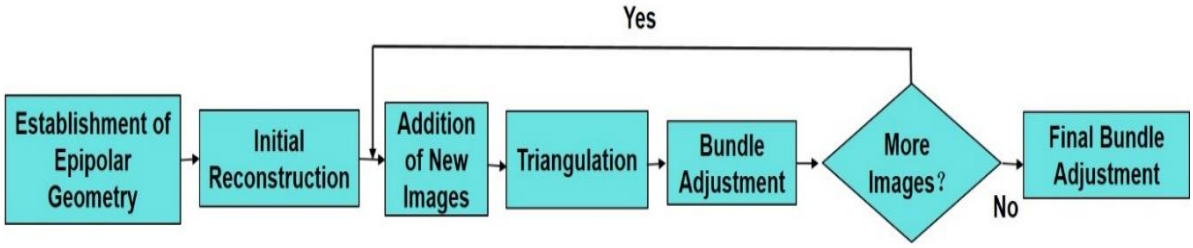


Figure 2.1.: Overview of the workflow of incremental image orientation methods

Over the last twenty years, various incremental methods were proposed, all following the workflow shown in Figure 2.1. One of the most well-known works was presented by Pollefeys et al. [2004], in which a hand-held camera was incrementally oriented using self-calibration. Similar to Figure 2.1, the set of images for initial reconstruction is selected by taking the number of correspondences and the length of the baseline into account. New images were added based on projective resection and intersection (triangulation). The results were refined by bundle adjustment, together with optimizing the focal length. Typically, the automatic orientation of uncalibrated images cannot guarantee to deliver results of high accuracy [Remondino and EI-Hakim, 2006]. Thus, in practice, especially in the field of photogrammetry, cameras are typically well calibrated in advance in a controlled environment, e.g., using a planar chess board [Zhang, 2000], a 3D test field [Lumann et al., 2014] or a control field generated on a liquid crystal display (LCD) [Zhan, 2006].

In the last two decades, many notable works have appeared in the computer vision community, among which four well-known studies are reviewed. To the best of the author's knowledge, Snavely et al. [2006] was the first trial to cope with images from Internet sources. In their work they proposed the popular SfM system *Photo Tourism* which allows to reconstruct large scenes, such as famous tourist-spots, using an incremental image orientation approach. As Internet photos are typically characterized as unordered and highly redundant, Snavely et al. [2008] extended their previous work [Snavely et al., 2006] by conducting incremental image orientation along a skeletal graph. The basic assumption is that a small subset of images may already be sufficient to represent most of the information of a scene. Consequently, such subsets of images and their pairwise epipolar geometries are used to build the skeletal graph. Inspired by these two works, researchers further start to think about the feasibility of reconstructing city sized scenes. For this purpose, Agarwal et al. [2009] presented a system which utilizes multiple

computers as a computation cloud to deal with extremely large collections of photographs from the Internet, such as the city of *Rome*. In their work the distribution and parallelization of huge processing tasks including feature extraction, image matching, pairwise epipolar geometry verification and incremental image orientation were investigated and described in detail. Only one year later, Frahm et al. [2010] explored the possibility of reconstructing the city of Rome on just a single computer within one day, calling the approach "Building Rome on a Cloudless day". They first classify images into clusters using appearance-based features which is carried out on the GPU to increase the computational efficiency. Incremental image orientation using skeletal graphs is then performed on each cluster individually.

More recently, literature published on the task of incremental image orientation mainly concentrated on single steps of the workflow illustrated in Figure 2.1. Wu [2013], for example, improved incremental image orientation by proposing a Re-triangulation method to reduce the accumulated error. It is assumed that the initially estimated poses and also the poses obtained by bundle adjustment may not be accurate enough, while some correct feature matches may fail to be triangulated. Such failed feature matches were re-triangulated in the recursive scheme. Wang et al. [2018b], on the other hand, proposed a new image orientation method which belongs to the step "Addition of New Images": For every newly added image, the corresponding rotation matrix was estimated by the relative rotations between the newly added image and the already oriented images. The translation was solved subsequently, using a linear equation system. The work by Schönberger and Frahm [2016] has received wide attentions because of its contributions to almost all steps of the workflow shown in Figure 2.1 and the corresponding released package[1]. In the first two steps, they estimate various stereo models via fundamental matrix, homography matrix or essential matrix (if the interior orientation parameters are provided). Based on the ratio of inliers using different models, valid image pairs are kept and used to compute an initial reconstruction. In order to add new images, the next best view is then chosen according to both, the number of visible tie points and their distribution in each candidate image. To achieve a robust and efficient triangulation, an adaptive RANSAC technique is adopted using the DLT method [Abdel-Aziz and Karaa, 1971; Hartley and Zisserman, 2003], while two constraints with respect to the triangulation angle and the positivity of the depth are checked. After the addition of images and triangulation, bundle adjustment is conducted using a strategy which perform local bundle adjustment on the set of most-connected images after each image registration and a global bundle adjustment only after a certain number of images has been added. It is also worth mentioning that SLAM [Mur-Artal et al., 2015] is actually a special case of incremental image orientation. The main differences are that only selected keyframes are used for generating tie points and loop closure is detected online.

2.3.2 Hierarchical image orientation

Almost at the same time as incremental image orientation was developed, the idea of hierarchical image orientation started to capture the attention of researchers. The basic idea is

[1] More details can be found at https://colmap.github.io/

to split the images into smaller overlapping subsets, while each subset is solved individually before being ultimately merged into one complete block. Hierarchical sub-sampling is pioneered by Fitzgibbon and Zisserman [1998], using a balanced tree of triplets over a video sequence. The solved triplets are merged by minimizing a cost function with respect to either the error of aligning corresponding tie-points or the reprojection error in image space. This approach is subsequently improved by Nistér [2000], adding heuristics to suppress redundant frames and for triplet selection.

While the aforementioned hierarchical methods work under the assumption of sequential frames, Havlena et al. [2009] presented a new efficient technique for unordered datasets which includes three consecutive steps: First, an image similarity matrix was computed using the *tf-idf* idea; Second, atomic 3D models reconstructed from triplets were used as seeds for a hierarchical merge to form a larger 3D model. Finally, in case that single images existed which were not connected, these were glued to the best partial reconstruction based on the number of visible tie points. Another hierarchical method using triplets was presented by Mayer [2014] and a corresponding extended version by Michelini and Mayer [2020]. Starting from triplets, the merging procedure was performed hierarchically with two images between shared two connected triplets. The efficiency of the merging procedure was improved by randomly deleting shares of tie points. Comparing different strategies of deleting tie points, the authors concluded that the random deletion of points typically generates reliable and precise results.

The approach of Faranzena et al. [2009] is applicable to unordered sets of images as well. In this approach, the images are structured into a hierarchical cluster tree, while the computation of image orientation parameters is conducted by following this tree from the leaves to the root. Consequently, the task was split into smaller instances which were then separately solved and combined subsequently. The images were grouped by agglomerative clustering which produces a hierarchical, binary cluster tree. The simple linkage [Duda and Hart, 1973, pp 98-105], denoted as the distance between two clusters, is measured as the distance of the two closest objects within the corresponding clusters. One year later, Gherardi et al [2010] proposed an approach which considers several close clusters. A good compromise was achieved using the five closest clusters which, according to the authors, provides a good balance between the number of clusters and the tree height. Toldo et al. [2015] extended these two studies working with uncalibrated images and further demonstrated the effectiveness of the corresponding hierarchical strategy.

2.4 Global image orientation

In the previous section, some representative works of incremental and hierarchical image orientation have been reviewed. In this section, the focus is on global image orientation methods. Instead of sequentially solving the set of orientation parameters, based on the relative orientation information obtained from the establishment of epipolar geometry, global image orientation methods aim at computing all the available images' exterior orientation parameters

simultaneously. This approach indicates the meaning of the term global in the context of global image orientation. Furthermore, the term global also refers to a concept in which the whole set of relative orientations is taken into account jointly, while rotations and translations are estimated separately. Consequently, on the one hand global image orientation ensures that errors are not accumulated but distributed more evenly over all orientation parameters (e.g., closed loops in a dataset are inherently taken care of, because all redundant relative orientations are used). On the other hand, global image orientation methods are more sensitive to outliers in relative orientations.

In the following, some important contributions in the fields of outlier detection in relative orientations, global rotation estimation and global translation estimation are reviewed.

2.4.1 Outlier detection in relative orientation

Many related works have focused on detecting blunders in ROs, i.e., image pairs with incorrect relative orientation parameters. A conventional way based on RANSAC is to use the epipolar geometry constraint after image matching, in which the essential or the fundamental matrix is estimated using appropriate algorithms [Hartley and Zisserman, 2004; Nistér, 2004]. The ROs are only considered to be correct if a minimum number of point pairs conform with the model of central perspective. Although many wrong ROs can be eliminated in this way, non-overlapping pairs may still exist resulting from repetitive structure (RS) and very short baselines (VSB). Many works try to detect these errors. Here, they are divided into three categories: missing correspondences analysis, loop consistency constraint analysis and other methods.

Missing correspondences analysis. [Zach et al., 2008] first employed the so-called missing correspondences among an image triplet to infer incorrect ROs. The main idea is that if a substantial portion of correspondences between two images from the triplet cannot be observed by the third image, then the relative orientation between the two images is potentially incorrect. The authors used a Bayesian framework for all image triplets to check the correctness of the corresponding image pairs. [Roberts et al., 2011] improved this idea by verifying incorrect ROs via an expectation-minimization method which integrates the cues of missing correspondences and timestamp information. However, this information is not available in general, e.g., for unordered images the acquisition sequence is unknown. [Jiang et al., 2012] extended the missing correspondences idea by minimizing a target function which considers the number of missing correspondences across the entire reconstruction instead of the triplets. Specifically, a spanning tree is first built and then problematic ROs are iteratively detected in a greedy way. As a consequence, the method may get stuck in a local minimum.

Loop consistency constraint analysis. [Zach et al, 2010] developed a method which adopts the loop consistency constraint to infer the validity of ROs. They first generate cycles in the overlap graph; the relative rotations are then concatenated within each cycle. If all ROs in the cycle are correct an identity mapping should be obtained as a result. Potential errors are identified using a Bayesian network. [Reich et al., 2017] presented a sequential graph optimization method to eliminate incorrect relative rotations. Both [Reich et al., 2017] and [Zach et al., 2010] need a

long processing time when dealing with a large image dataset where all the relative rotations need to be considered. [Shen et al., 2016] presented a graph-based method, where a minimum spanning tree is incrementally expanded by checking the loop consistency within a triplet until all available images are included in the minimum spanning tree.

Other methods. [Wilson and Snavely, 2014] proposed a 1DSfM approach. Their basic idea is to project the 3D relative translations into different 1D direction vectors. They then used a kernel density estimator to sample these directions and showed that typically, wrong ROs clearly stand out in some directions of the 1D vectors. However, as the authors write, their method fails in the presence of repetitive structure. To address this issue, [Wang et al., 2018a] presented a hierarchical RO selection method for repetitive structure. They first built a minimum spanning tree (MST), and then used a hierarchical scheme for RO selection. The method only selects validated ROs along the MST and thus may break up a block of images, while image pairs with very short baselines are not dealt with. To solve for artefacts caused by repetitive structure, [Cohen et al., 2012] considered various symmetrical structures using geometric and appearance cues, to refine their bundle adjustment process. [Heinly et al., 2014] presented a post-processing step in the sense that the result of SfM is the input of their method. They split the overlap graph into subsets and use conflicting correspondences to identify repetitive structure. The subsets of the overlap graph which are free of conflicts are then merged into a correct reconstruction.

2.4.2 Global rotation estimation

Many approaches have been proposed to solve the problem of global rotation estimation from pairwise relative rotations. In general, they can be categorized into direct solutions and coarse to fine solutions.

Direct solution. The basic idea is to estimate global rotation by using all available relative rotations. As one of the pioneered works, [Govindu, 2001] parameterized rotations as quaternions and determined the global rotation parameters by a constrained least-squares optimization. More specifically, the residuals between the measurements of relative rotations and the corresponding estimated global rotations were minimized under the assumption of a Gaussian distribution for the relative rotations' uncertainty. However, the fact that a quaternion has unit length to unambiguously describe a rotation (up to the sign) and blunders in relative rotations are not taken care of. [Govindu, 2006] developed a robust rotation averaging method, where the Lie algebraic approach was exploited to provide a closed-form solution in a more flexible, fast, and accurate manner. In addition, outliers are handled robustly by applying a RANSAC technique: independent minimum spanning trees (MSTs) are used for extracting global rotations and the corresponding distances between the estimated global rotations to the relative rotations are determined along the MSTs. After a number of trials, the solution with the largest number of inliers is refined and optimized using Lie algebra. [Olsson and Enqvist, 2011] presented a strategy similar to [Govindu, 2006], however, the improvement lies in adding a weight to each relative rotation, which is proportional to the total number of matching points. Other methods suggest a linear solution by relaxing the constraints on rotation parameters: [Martinec and

Pajdla, 2007] and [Moulon et al., 2013] investigated a quaternion representation for solving image rotations and found that there was no satisfactory way to solve a linear system due to the required quadratic constraints of unit length quaternions. Thus, they alternatively first solved a homogeneous equation system using SVD (singular value decomposition) and then projected each approximate rotation estimation to the closest $SO(3)$ space (in terms of Frobenius norm). While using SVD, they enforced the matrix to satisfy the orthonormality constraint required from rotation matrices. [Arie-Nachimson et al., 2012] presented a similar spectral decomposition method which typically performs better on datasets with large ratio of outliers compared to the method of [Martinec and Pajdla, 2007]. This superiority is attributed to their formulation of the rotation averaging problem in a form of semidefinite program (SDP) relaxation. In detail, the problem is cast into a trace maximization for a product of symmetric matrices composed of relative and global rotation matrices. A constraint causing a tighter convex relaxation of the optimization problem is integrated, thus, a solution which is less sensitive to noise and mismatches is explicitly promoted.

Coarse to fine solution. Initial coarse rotations are estimated first using only a part of input relative rotations, before the following refinements are conducted by using all redundant relative rotations. Hartley et al. [2011] built a minimum spanning tree and computed the initialization by propagating relative rotations starting from the root of this tree. The estimated initial values were iteratively optimized using the Weiszfeld algorithm [Weiszfeld, 1937] (in French) or [Weiszfeld and Plastria, 2009]. They also compared the cost functions based on L1 and L2 norm, and showed that the L1 norm was markedly better with respect to robustness. DISCO [Crandall et al., 2011] adopted a hybrid discrete-continuous optimization scheme. The authors treated the initialization problem as a classification task for which the $SO(3)$ space is divided into 1000 independent labels. Based on a Markov random field formulation of constraints between relative rotations, discrete belief propagation was used to determine a specific label for each image. Finally, a non-linear least squares optimization is used to refine the initial rotations. It is demonstrated, in the experiments of Chatterjee and Govindu [2013], that both Crandall et al. [2011] and Hartley et al. [2011] are limited by the computational efficiency and the requirement of a scalable robust rotation estimation scheme. Therefore, a notable two-stage approach was proposed in Chatterjee and Govindu [2013]: They first calculated rather robust initial values by averaging the L1 norm on the Lie algebra (vector space). The solution was refined by iterative reweighted least squares, implemented by a Huber-like loss function. In particular, the corresponding weights were dynamically determined based on the residuals on the Lie-algebraic vector space. [Reich and Heipke 2015, 2016; Reich et al. 2017] improved the approach of Chatterjee and Govindu [2013] by employing a convex relaxed semidefinite program to obtain a more robust initial solution, before a refinement scheme is performed using a sigmoidal weighting function [Krarup et al., 1980].

A more comprehensive review of rotation averaging was published by Hartley et al. [2013] which can be understood as a tutorial, discussing, for example, different distances of rotation including geodesic distance, angular distance and quaternion distance as well as investigating different loss function such as L1 and L2 norm. Wilson et al. [2016] made a further contribution

on rotation averaging. In their study, two heuristics were suggested: First, the rotational gauge ambiguity is proposed. Typically, the first image's rotation matrix is assumed to be an identity matrix. To achieve a higher local convexity, it is beneficial to choose a starting image that is connected to many other images instead of using one from the periphery of the block. Second, the analysis of the normalized graph Laplacian [Luxburg, 2007] is suggested to decide whether it is a well-connected graph (which is easier to solve) or a larger and noisier graph (which may be hard to solve). This motivates a multi-stage method to solve larger, less connected problems by first addressing smaller, simpler and well-connected subgraphs.

2.4.3 Global translation estimation

Analogous to global rotation estimation, global translation estimation is normally formulated as computing global translation parameters from the relative translations of all available image pairs. However, different from global rotation estimation, global translation estimation cannot be conducted directly from relative translations, because a relative translation only contains the normalized translation vector between the projection centers of two images. Without individual scale values relating all image pairs to a global unified scale, global translation parameters cannot be directly calculated. Two lines of research were developed to solve this issue, the first one has a focus on only using relative translations, the second one lies on taking both, relative translations and information of tie points, into consideration.

Solution with relative translations only. The basic idea is inspired by the fact that the estimated relative translation should be parallel to the translation vector computed from the corresponding pairwise global translations (the discrepancy value is denoted as unparalleled error). Govindu [2001] proposed a linear framework based on every pairwise constraint indicated by the corresponding unparalleled error, in which the size of the error of each individual constraint varies a lot due to the various baseline lengths between different image pairs. To further refine the solution, they conducted an iterative weighted adjustment with the goal of unifying the weight of each individual constraint, in fact, after convergence these unified weights are the global unified scales related to corresponding image pairs. However, this method does not cope with gross errors, therefore, it is very sensitive to outliers. Brand et al. [2004] estimated the translations by minimizing the sum of all squared unparalleled errors. To fix the global scale and to prevent degeneracy, they add two additional constraints requiring the sum of translations and squared translations to be equal to zero and ones, respectively. Similar to these two methods, Wilson and Snavely [2014] used the sum of squared chordal distances which leads to a non-linear optimization problem. The relevant convergence properties were analyzed, and a good solution can consistently be found once the majority of outliers have been removed. Ozyesil et al. [2015] improved the method of Brand et al. [2004] by integrating a non-convex constraint that requires the estimated translations to satisfy a maximal proximity condition. This condition requires that two images should have a minimum distance from each other. Finally, a semidefinite relaxation formulation is used to remove the non-convex constraint. Moulon et al. [2013] modified the unparalleled error via embedding the unknown scale factors which reconcile different relative translations into a global coordinate frame. Their approach consists

of an optimization under the L_∞ norm which can be formalized in the form of a linear program. Arrigoni et al. [2016] extracted a number of so-called basis circuits, in which the local unified scales can be decomposed by utilizing the geometric characteristics of these circuits. Global unified scale factors can be estimated by propagating the scales of each basis circuit, before the global translations are estimated using the relative translations and the corresponding global scales. Zhuang et al. [2018] discussed the influence of the length of image baselines on estimating global translations. As the location precision is sensitive to baselines of different lengths, they advocated an objective function using an angular error which is independent of the baseline length.

All the above-mentioned approaches are demonstrated to be able to generate accurate results if the locations of the image projection centers are evenly distributed in object space. Nevertheless, they are all invalid when dealing with images that are taken along a straight line. In other words, they fail to deal with images whose projection centers are collinear. To solve this degenerate case and to improve the robustness, approaches which include tie points have been investigated.

Solution combining relative translations and tie points. Next to rotation estimation, Crandall et al. [2011] also described a hybrid discrete-continuous optimization scheme to estimate global translations. In this scheme, the 3D space was split into a certain number of subspaces denoted as labels and the number and the size of these labels were determined according to the information from the embedded geotags of the images. Based on a Markov random field formulation of constraints between relative translations and camera-point direction vectors, discrete belief propagation was used to determine a specific label for each image. A non-linear least square optimization is then performed to refine the initial translations and rotations. However, this approach is applicable only if some prior knowledge (e.g., geotags) is available. Arie-Nachimson et al. [2013] derived an expression for the essential matrix in terms of a global coordinate system. In particular, the essential matrix was rewritten using the already estimated global rotations and the global translations to be solved. A linear equation system was then set up, based on the epipolar geometric relationship between essential matrix and correspondences. Jiang et al. [2013] considered this problem from a different view. They used image triplets and the corresponding geometric relationship constrained by common object points to set up a linear equation system for determining the translation parameters. As a consequence of requiring triplets rather than pairs, some images may not be included in the resulting block, and their pose is then not recovered. Reich and Heipke [2016] improved the work of Jiang et al. [2013] by introducing multi-ray points, i.e. points visible in multiple images. Nevertheless, images may remain unconnected. Cui et al. [2015] chose multi-ray tie points, whose corresponding 2D image coordinates were refined as inliers by epipolar geometry verification. From the corresponding 2D image coordinates, a linear equation system was built to determine all translation parameters simultaneously using the L1 norm. Obviously, the results are affected by the choice of tie points. Another work from Cui and Tan [2015] also computed the global unified scales first, as Arrigoni et al. [2016], but the scales are unified using the depth information from each individual local spatial intersection. Afterwards, a linear equation system

determined by the solved global rotations, relative translations and their scale factors, is built to solve for the global translations. By using collinearity equations and the information of tie points, Wang et al. [2019a] propose a linear global method. Given the global rotation and tie point information, they first selected some robust tie points that can connect all available images into the same photogrammetric block. Then, the translation parameters and selected 3D tie point coordinates are solved simultaneously. But, as the number of images increases, so does the number of unknown tie points, which brings much more computational burden for the linear global method.

2.5 Alternative solutions for image orientation

Some ideas were published to avoid having to compute rotation and translation separately. Bourmaud et al. [2014] derived the image pose parameters as a Lie group *SE(3)*. The authors proposed a generative model based on the formulation of a concentrated Gaussian distribution on the matrix Lie group and solved an iterated extended Kalman filter on that group to compute the elements of *SE(3)*. Kasten et al. [2019a] proposed a method to globally recover the projection matrix of each image by using fundamental matrices of image pairs. However, as the projection matrix yields a projective reconstruction, information on interior orientation parameters cannot be introduced. Later, the authors extended their work: Exploring the algebraic characterizations of essential matrices, they introduced a method to simultaneously solve for rotation and translation of each image from essential matrices [Kasten et al., 2019b], a corresponding degenerate case occur if all images' projection centers are (or nearly) collinear. Recently, Geifman et al. [2020] further characterized the algebraic characterization of essential and fundamental matrices in collinear image translations settings. They also suggested a practical solution for treating tie points as cameras to remove near-collinearity degenerations. However, this results in a much larger optimization problem that needs to be solved, leading to a large runtime. Furthermore, robustly choosing tie points for the optimization problem is a challenging task by itself. Cui et al. [2017] described a hybrid method consisting of a global and incremental strategy, in which rotations were determined by global rotation estimation [Chatterjee and Govindu, 2013] and translations were solved in an incremental manner. Later, Cui et al. [2019] improved this method with respect to robustness and time efficiency. Spurious image pairs are detected in the process of global rotation estimation, the corresponding image matches are eliminated and a subset of well-conditioned tie points are selected to accelerate the most time-consuming final procedure - bundle adjustment. To overcome the degenerate collinear case, Wang et al. [2021] proposed a hybrid global image orientation by extending the work of [Kasten et al., 2019b]. More specifically, an efficient method for extracting an optimal minimum cover connected image triplet set (OMCTS) is proposed, this OMCTS makes all available images included by a minimum number of connected triplets, as well as all of those selected triplets, satisfy the constraint that the three corresponding relative orientations are as compatible as possible to each other, after that, in the OMCTS the collinear triplet (invalid in [Kasten et al., 2019b]) and non-collinear triplet are solved separately, finally, all image

orientations in a common coordinate system are estimated by traversing solved connected triplets using a similarity transformation.

2.6 Discussion

The core steps of image orientation are reviewed and for each step, the corresponding state-of-the-art works are studied. This section provides a brief insight of the limitations and open questions in the context of the research objectives of this work (namely, efficient image matching, outlier detection in relative orientations and estimation of image orientation parameters). Based on these open issues, corresponding methods are proposed in the following chapters to tackle the identified research gap.

Image matching

To reduce the effort that is required for image matching, three strategies are investigated: first, reducing the number of features per image for improving the time efficiency of image matching. One way for this reduction is to manually vary some inherent thresholds (e.g., DoG [Vedaldi and Fulkerson, 2008] or the scale value [Wu, 2013]) when generating features. Another way is via training [Hartmann et al., 2014], in which only the features that are predicted to be suitable for matching are kept. Although less feature can ease the computational burden, this can result in the undesired fact that *less projection rays are constructed in the photogrammetric block which is not advantageous for the solution of image orientation* (especially, for images with week connection to the block). Second, reducing the number of image pairs. Several studies [Nistér and Stewenius, 2006; Mur-Artal et al., 2015; Zhan et al., 2015;] are explained based on the idea of "Bag of Word", which is widely used in image retrieval. In addition, to further reduce the computation of "Bag of Word", some global features, e.g., deep convolution features [Wan et al., 2018; Zhan et al., 2018] and gist feature [Frahm et al., 2010], are employed instead of local hand-crafted feature (e.g., SIFT). Two limitations are found in this strategy: *1) ambiguous results can be generated if any two 'words' are similar to each other. 2) When using global features, some additional efforts have to be taken to generate local features that can be used for subsequent image orientation*.

To address the mentioned research gaps, this thesis develops a fast method for detecting mutually overlapping pairs, in which several random k-d trees are constructed based on partial extracted SIFT features per image, and correspondences are then generated only using the detected overlapping pairs with all extracted features. As a consequence, similar to the strategy of reducing the number of features per image, partial features are employed when building random k-d trees, whereas, the number of projection rays is supposed to be of sufficiently high redundancy, as all the extracted features are used for generating conjugate points. On the other hand, to avoid the ambiguity of the indexing structure, the random k-d trees are constructed in a way that makes them as independent to each other as possible. Lastly, SIFT features are

inherently used in this work, thus, no additional efforts are needed to dealing with image features.

Image orientation

To solve the problem of image orientation, three common strategies that frequently appear in related publications are reviewed: ***incremental***, ***hierarchical*** and ***global*** image orientation methods. In the previous sections, some representative works of ***incremental*** and ***hierarchical*** image orientation have been summarized, however, both of these two strategies are limited by low time efficiency due to the repetitive usage of bundle adjustment. To cope with this problem, in this thesis, the ***global*** image orientation method which only needs the final bundle adjustment is employed. In addition, its process of estimating image orientation parameters ensures that random uncertainties are minimized with respect to all orientation parameters, and do not accumulate to systematic effects such as bias or drift. The disadvantage of ***global*** methods is, on the other hand, that outliers in relative orientation have larger negative effect on the results.

To detect outliers in relative orientation, three categories of works are discussed, namely, missing correspondences analysis, loop consistency constraint analysis and other methods. According to the corresponding descriptions, they only show good performance on detecting specific type of RO outliers, in detail, *missing correspondences analysis can deal with RO outlier due to repetitive structure (RS) but fail on RO outliers due to very short baseline (VSB). Loop consistency constraint analysis and some other methods [Wilson and Smavely, 2014] are able to detect the RO outliers due to inappropriate baseline and noisy observations but have difficulties on detecting RS RO outliers*. To improve the robustness of the proposed global image orientation method and to overcome various RO outliers, in this thesis, a general method via checking the triplet compatibility is first presented to deal with RO outliers due to noise observations, and a combined ROs robustified method is then developed to deal with RO outliers due to RS and inappropriate baselines.

Similar to most global methods, in this thesis, the two-step global image orientation consisting of global rotation estimation and global translation estimation is adopted. As ample research has been published on global rotation averaging in recent years and this problem can be considered to be nearly solved, an existing method is used in this thesis (the one suggested by Chatterjee and Govindu [2013] which can be seen as a state-of-the-art method for a robust, accurate and efficient computation of global rotations, used in several recent publications on global image orientation [Wilson and Snavely, 2014; Ozyesil and Singer, 2015; Cui et al., 2015]).

Global translation estimation, in contrast, is still receiving a lot of attention from researchers. Among them, two lines of works are studied: solution with relative translations only and solutions combining both relative translation and tie points. In the first line, many relevant approaches are proposed by minimizing the unparalleled error between the relative translations and the translation vector calculated by the corresponding global translations. *However, they are all invalid when all the input images' projection centers are collinear*. In the second line, to cope with the invalid case existing in the first line, tie points are integrated in some way, e.g.,

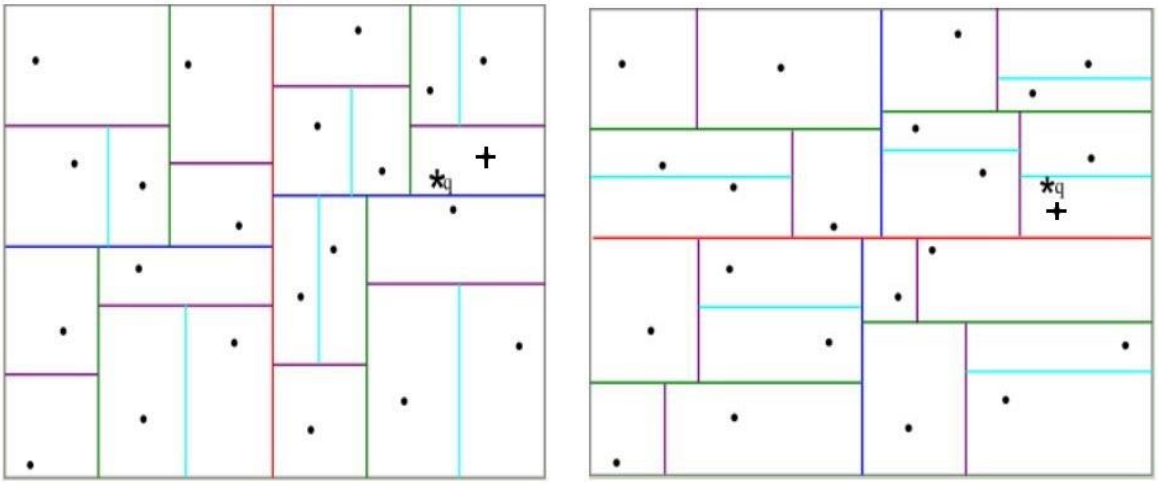
the epipolar constraint was reformulated by using the tie point information [Nachimson et al., 2013] or the same tie point viewed by different images should stay in the same position [Jiang et al., 2013; Cui et al., 2015; Reich and Heipke, 2016; Wang et al., 2019b]. All these methods suffer from the selection of tie points, specifically, *a sufficient amount of tie points have to be selected to connect all images which results in a very large linear optimization problem and outliers exist in the selected tie points which leads to imprecise solution*. Following the second line, a new global translation estimation method is presented, in which the invalid case of the first line does not exist anymore. In addition, unlike most of the investigated methods belonging to the second line, in this work, only robust tie points within triplets of images together with the relative translations are used to estimate the global translation parameters.

3 Preprocessing

3.1 Time efficient image matching based on a random k-d forest

In this section, an approach is described for determining overlapping image pairs fast, i.e. image pairs forming a stereoscopic model, in a set of unordered images, cf. [Wang et al., 2017]. The suggested approach is feature-based, and relies on a random k-d forest made up of several independent k-d trees for nearest neighbor search. An algorithm is proposed to compute the degree of similarity of images based on these nearest neighbors.

3.1.1 Construction of the random k-d forest



(a) (b)

*Figure. 3.1: Two different k-d tree structures [Muja and Lowe, 2014]. For the same dataset, two random k-d trees are shown, the splitting hyperplanes in (a) and (b) are different and independent from each other. The retrieved nearest neighbor of **q** by these two k-d trees is indicated by +, in (b) the nearest neighbor (with shortest distance to **q**) is correctly found, whereas for (a) the result is not correct.*

K-dimensional or k-d trees are binary search trees and provide well-known solutions for conducting nearest neighbor search with both, high efficiency and precision, especially for low dimensional data [Robinson, 1984]. Given a k-d tree as shown in Figure 3.1(a), the nearest neighbor of a query feature ***q*** would be determined to be the one falling into the same cell, which, as can be seen from the figure, may be incorrect. One classical way to solve this problem is to backtrack to the parent nodes step by step and compare the distance between ***q*** and the features

from the parent nodes, until the correct nearest neighbor is found. However, in higher dimensional data, a large number of nodes may have to be traversed, and when faced with a large amount of data, finding the nearest neighbors in an exact sense requires much effort for backtracking, which reduces the retrieval efficiency. To solve this problem, approximate methods have been developed. As an example, Arya et al. [1998] proposed the so-called priority search algorithm. Another efficient method is to provide a set of different k-d trees; Figure 3.1(b) shows a second one. In this case, the true nearest neighbor does fall just into the cell of ***q***. Thus, using a set of k-d-trees, called a random k-d forest, improves the probability that the query feature and the nearest neighbor fall into the same cell in at least one tree, and only this cell needs to be checked.

In line with existing literatures (see section 2.2), based on a random k-d forest, a method for finding overlapping image pairs is presented. The input for each tree is the set of features extracted from all images. SIFT features are applied with the corresponding descriptor of 128 entries normalized to a length of 1 [Lowe, 2004], Only a subset of all extracted features is used (the other features are stored for later usage) to increase the time efficiency of the proposed method and to make the computational resources more feasible for larger datasets. In this thesis, 60 percent of the extracted features per image are used if the number of image is smaller than 500, whereas, the percentage is reduced to 50 if the number of image ranges from 500 to 1500, and it keeps decreasing to 40 percent if the number of images is higher than 1500.

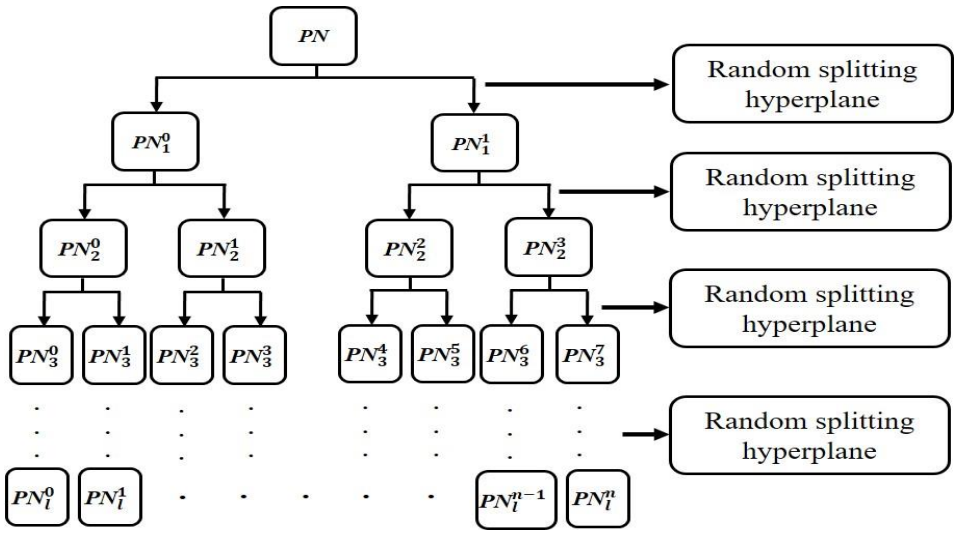


Figure. 3.2: The generation of one k-d tree.

To build the random k-d forest from the selected features, the following rule is suggested: the k-d trees of the forest should be as independent from each other as possible, i.e., each k-d tree should have a different tree structure. The detailed construction procedure is illustrated in Figure 3.2, the parent node ***PN*** of the tree is presented with the SIFT descriptors of all input images. The feature space is recursively split into two regions corresponding to child nodes by a hyperplane, until the child nodes contain at most one descriptor. When building k-d trees, one conventional way is that the entry of the feature vector with the largest variance is chosen for splitting. In order to make the trees more independent, the splitting plane is randomly selected from a range of entries

with relatively large variances, and the hyperplane selection is randomly conducted anew for every splitting operation. In this thesis, *n_tr* k-d trees are built using the same rule.

After generating the random k-d forest, priority search according to Arya et al. [1998] is applied in each of these k-d trees. Therefore, for a query feature, each tree can contribute nearest neighbors. Among these nearest candidates, those fulfilling the following three constraints are selected for further processing:

a) the candidate is from an image *j* different from the image *i* of the query feature;
b) the scalar product d_{ij} of the query and the candidate feature vectors is above a threshold *d* (note that all features are normalized to unit length);
c) if there are multiple candidates from the same image, the one having the largest scaler product d_{ij} is kept.

3.1.2 Determination of overlapping image pairs

According to the algorithm described in section 3.1.1, the random k-d forest is built. The ID of each SIFT descriptor is also recorded so that the information of which image the descriptor came from is preserved. In this way, it can be ensured that for a query feature only features from other images are considered. The nearest neighbors of each feature from the *i*-th image (i =1,2, 3, …, N) are retrieved by traversing the random k-d forest, where N is the number of images. The number of the resulting neighbors per image pair *ij* is called P_{ij}, where *j* is the image ID of the *j*-th image. The larger the value of P_{ij}, the more potential matches between the *i*-th and *j*-th image exist, and the more likely it is that the two images overlap.

Equation (3.1) is proposed to calculate the degree of similarity S_{ij} between two images *i* and *j* as a function of P_{ij} and D_{ij}, where D_{ij} is the average of all values d_{ij} involving correspondences of features from image *i* and *j* (note that to determine a relative orientation between two calibrated images, at least 5 corresponding features are necessary, i.e. P_{ij} must be at least 5. S_{ij} is set to -1, if P_{ij} is less than 5). The more similar the *i*-th and *j*-th images are, the larger the value of S_{ij} is.

$$S_{ij} = e^{D_{ij}} \cdot \log_{10} P_{ij} \tag{3.1}$$

Equation (3.1) is a heuristic measure of image similarity. To motivate its construction, it is argued as follows: P_{ij} and D_{ij} should have relevant influences on the results, D_{ij} is always smaller than 1, whereas P_{ij} might take on a very large value depending on the number of feature correspondences (note that the value of P_{ij} can vary considerably when dealing with images of different resolution). P_{ij} can thus easily dominate a similarity measurement which is not desirable. Therefore, the influence of P_{ij} should be decreased and that of D_{ij} increased; this goal is achieved by using the logarithm for P_{ij} and the exponential function for D_{ij}.

The similarity degree values S_{ij} are calculated according to equation (3.1) for each pair of images *i, j*. For each image *i,* these similarity degree values are sorted by size in descending order, and image pairs that have the *a*% largest S_{ij} scores are chosen as potential overlapping pairs for image *i*.

Algorithm 3.1 Clustering images and discarding single images

Input Symmetric $N*N$ matrix $\boldsymbol{Q}$.

Output p symmetric matrices $\boldsymbol{Q}_i$, where p is the number of clusters.

I. Initiate a new symmetric $N*N$ matrix, called adjacency matrix $\boldsymbol{AM}$. If $\boldsymbol{Q}_{ij} > cp_{min}$, set $\boldsymbol{AM}_{ij} = 1$, otherwise, $\boldsymbol{AM}_{ij} = 0$.

II. Initiate a new container V :={-1} of size N, an integer vt =0 and integer $t = 0$
do
{
1. Create a new empty container IC_t. If $\boldsymbol{AM}_{ij} = 1$, add i and j into container IC_t , and set $V_i = 0$, $V_j = 0$.
2. Traverse the current container IC_t, add the images' ID into container IC_t whose corresponding $\boldsymbol{AM}$ values are equal to 1.
3. Repeat 2 until the size of the current IC_t does not change any more.
4. Set vt equal to the number of 0 element of IC_t; $t = t+1$.
} while(vt != N)

6. Multiple containers (IC_t) are generated according to the number of iterations. The images which are classified into the same container IC_t belong to the same cluster.

III. From the containers IC_t and $\boldsymbol{AM}$, the image overlap $\boldsymbol{Q}_i$ can be determined in each cluster. $\boldsymbol{Q}_i$ represents the overlap result of the i-th cluster, only the largest $\boldsymbol{Q}_i$ is kept for subsequent processing.

3.1.3 Clustering images and discarding single images

Typically, large sets of unordered images, especially crowd-sourced images collected from websites, contain several smaller unconnected clusters, i.e. there are possibly no or not enough tie points between images from different clusters. Furthermore, sometimes single images exist which do not have any overlap with any of the other images. Based on the information from sections 3.1.1 and 3.1.2, it is easy to count the number of putative conjugate points for each image pair and to store these values in a symmetric $N*N$ matrix $\boldsymbol{Q}$. In order to obtain stable relative image orientations, the requirement that each potentially overlapping pair has a minimum of cp_{min} conjugate points is set up. Subsequently, an adjacency matrix $\boldsymbol{AM}$ is derived, where the entry at position (i, j) is 1 if the corresponding image pair fulfils this criterion, otherwise this entry is 0. $\boldsymbol{AM}$ is then recursively traversed to determine which images belong to which cluster. In this work, it is assumed that most images are a part of one and the same photogrammetric block. Consequently, the cluster with the largest number of images is investigated further, and all single images and all smaller clusters are deleted (see Algorithm 3.1 for more details).

3.1.4 Determination of relative orientation parameters

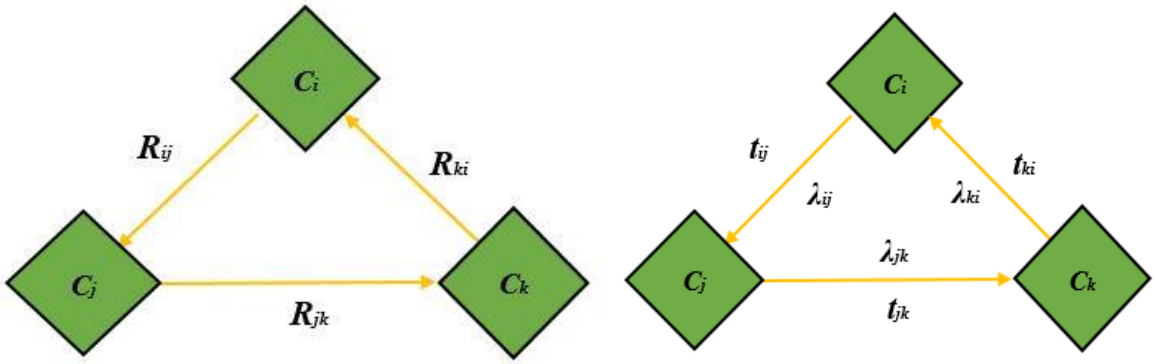
Having determined overlapping image pairs, the relative orientation parameters of all these pairs are computed. In this step, all extracted features are employed, i.e., those used to build the k-d

forest, together with the additional features not used in the overlap relationship determination. These additional features are matched based on the feature descriptors using pairwise image matching.

Relative orientation is represented by the essential matrix $\boldsymbol{E}$ if the corresponding interior orientations of the images are known, and its elements are obtained by using the five-point algorithm [Nistér, 2004]; otherwise, the fundamental matrix $\boldsymbol{F}$ is used. In both cases RANSAC is employed for blunder detection. Relative rotation and translation are then derived from $\boldsymbol{E}$ or $\boldsymbol{F}$. Image pairs, together with the correspondences and relative orientation parameters, with at least a pre-defined number (N_c) of conjugate points are kept, where this number must also account for more than b% of the number of correspondences used as input. From the relative orientation of remaining pairs, a viewgraph is constructed, in which images are indicated as nodes, while edges denote two images whose relative orientation is successfully derived.

3.2 Robustifying the ROs for robust global image orientation

Revisiting the objective of pursuing a robust way to solve image orientation, this section strives to improve the robustness of global image orientation and to deal with outliers in the input, e.g., relative orientations. While robust relative orientation estimation using the five-point algorithm combined with RANSAC [Fischler and Bolles, 1981; Nistér, 2004] can eliminate a certain number of outliers in the set of relative orientations, typically some wrong results remain after this step. In this section, a general method employing all triple-wise overlapping images is considered. Specifically, each triplet's compatibility regarding the inherent relative rotations and relative translations is checked. The basic assumption is that RO outliers typically result in a lower compatibility of related triplets (see below for the meaning of 'compatibility'). As Wang et al. [2019c] showed that some RO outliers stemming from repetitive structure (RS) and very short baseline (VSB) can pass the triplet compatibility check, methods to eliminate incorrect ROs which have resulted from RS and VSB are further investigated and two corresponding criteria that indicate the quality of ROs are presented. RS is detected based on counts of conjugate points of the various image pairs, while VSB is found by inspecting the intersection angles of the corresponding image rays. By investigating these two criteria, incorrect ROs are detected and eliminated using some empirical settings. As correct ROs of image pairs with a longer baseline nearly parallel to both viewing directions (BPVD) can be valuable, a method to identify and keep these BPVD ROs is also part of this section. In particular, the individual correspondences of BPVD ROs are analyzed via the cofactors of corresponding object points during triangulation, while deleting those with unreasonable values.



(a) triplet closure of relative rotations *(b) triplet closure of relative translations*

Figure 3.3: A closed loop from an image triplet.

3.2.1 Detecting and eliminating RO outliers by checking compatibility of triplets

A triplet is formed by three images that overlap each other. Two potential triplet closure discrepancies, in terms of relative rotations and translations as Figure 3.3 shows, are studied to specify the triplet compatibility, respectively. Note that all relevant triplets formed by three mutual overlapping images are investigated

Rotation outlier detection

Knowing one triplet's three relative rotations (i.e. $\boldsymbol{R_{ij}}$, $\boldsymbol{R_{jk}}$ and $\boldsymbol{R_{ki}}$, as Figure 3.3 (a) shows), the result of $\boldsymbol{R_{ij}R_{jk}R_{ki}}$ should in principal be equal to $\boldsymbol{I_{3\times3}}$. However, this condition is typically not perfectly fulfilled due to noise and outliers from relative rotations. Thus, it makes sense to indicate a triplet's rotational compatibility by $d_{\angle}(D_R) = \arccos((tr(\boldsymbol{R_{ij}R_{jk}R_{ki}}) - 1)/2)$, where *tr*(.) returns the trace value. To detect RO outliers, the corresponding rotational compatibility is used as follows: If for a certain triplet the result of $d_{\angle}(D_R)$ is smaller than a threshold ε_r, all three relative orientations are regarded as inliers, otherwise, they are considered as potential outliers. In addition, if a relative rotation is an outlier, the corresponding relative translation is also considered to be incorrect, as relative rotation and translation are not determined independently. This procedure is repeated for each triplet. Then, each relative rotation is examined separately: if all triplets the relative rotation of interest is part of show a value above the threshold, this relative orientation is considered to be an outlier, otherwise it is an inlier.

Translation outlier detection

Given the derived consistent scale factors within triplets (which are computed by equation (4.13) in section 4.3.2), similar to the spirit of using $d_{\angle}(D_R)$, RO outliers can be further eliminated using the constraint $\lambda_{ij}\boldsymbol{R_i^T t}_{ij} + \lambda_{jk}\boldsymbol{R_j^T t}_{jk} + \lambda_{ki}\boldsymbol{R_k^T t}_{ki} = \boldsymbol{0}$, illustrated in Figure 3.3(b). In this context, $\boldsymbol{t_{ij}}$, $\boldsymbol{t_{jk}}$ and $\boldsymbol{t_{ki}}$, are the relative translations within one triplet, $\boldsymbol{\lambda_{ij}}$, $\boldsymbol{\lambda_{jk}}$ and $\boldsymbol{\lambda_{ki}}$ are the consistent scale factors and $\boldsymbol{0}$ is 3-dimensional zero vector, $\boldsymbol{R_i}$, $\boldsymbol{R_j}$ and $\boldsymbol{R_k}$ are the corresponding computed global rotations of image *i*, *j* and *k*. This constraint can typically not be strictly fulfilled either, because of imprecise relative translations and scale factors. Here, the same idea is employed as before: If the triplet

translation compatibility measure (computed using the L_2 norm $d_{\angle}(D_t) = ||\lambda_{ij}\boldsymbol{R_i^T t}_{ij} + \lambda_{jk}\boldsymbol{R_j^T t}_{jk} + \lambda_{ki}\boldsymbol{R_k^T t}_{ki}||_2$) is smaller than a threshold ε_t, all three translations are regarded as inliers, otherwise, they are considered as potential outliers. Again, this procedure is repeated for each triplet. Then, the relative translation of each edge in the viewgraph is examined separately: if all triplets containing the investigated relative translations show a value above the threshold, this relative orientation is considered to be an outlier, otherwise it is an inlier.

Summary

After identifying outliers in the described way, the viewgraph is updated by deleting all edges which have been found to be outliers. While some outliers may still be present in the resulting viewgraph, the characteristic of the presented procedure is that most correct relative orientations will not be discarded, leading to a relatively dense connection of the graph. In addition, when eliminating RO outliers by checking triplet compatibility, it is worth to note that global rotation estimation is performed right after rotation outlier detection, whereas global translation estimation can only be performed after both rotation outlier detection and translation outlier detection have been carried out. This is due to the fact that for the proposed approach translation outlier detection requires that consistent scale factors within triplets are already derived.

3.2.2 Detecting and eliminating RO outliers due to repetitive structure

Repetitive structure is a characteristic of a single image and describes the fact that multiple regions of the image look similar. Typically, this is caused by a repetitive 3D structure in the scene (also explaining the naming repetitive structure instead of repetitive texture, as texture refers to the 2D image space). If two images depict a scene with 100% repetitive structure, even well-trained people cannot interactively distinguish real overlapping image pairs apart from non-overlapping ones due to RS. To distinguish identify RS ROs from all ROs, one normally takes advantage of non-repetitive structure in the images. Figure 3.4 shows an example with four image pairs. For these pairs, correspondences can be generated by image matching as the red, green and yellow points in Figure 3.4 show. Visually, it's easy to tell that image pair 1 is a pair with real overlap since it contains non-repetitive structure (see the red boxes). In contrast, image pairs 2, 3 and 4, which are non-overlapping, do not have such non-repetitive structure. In order to determine repetitive structure between two images, the following assumptions are made in this work:

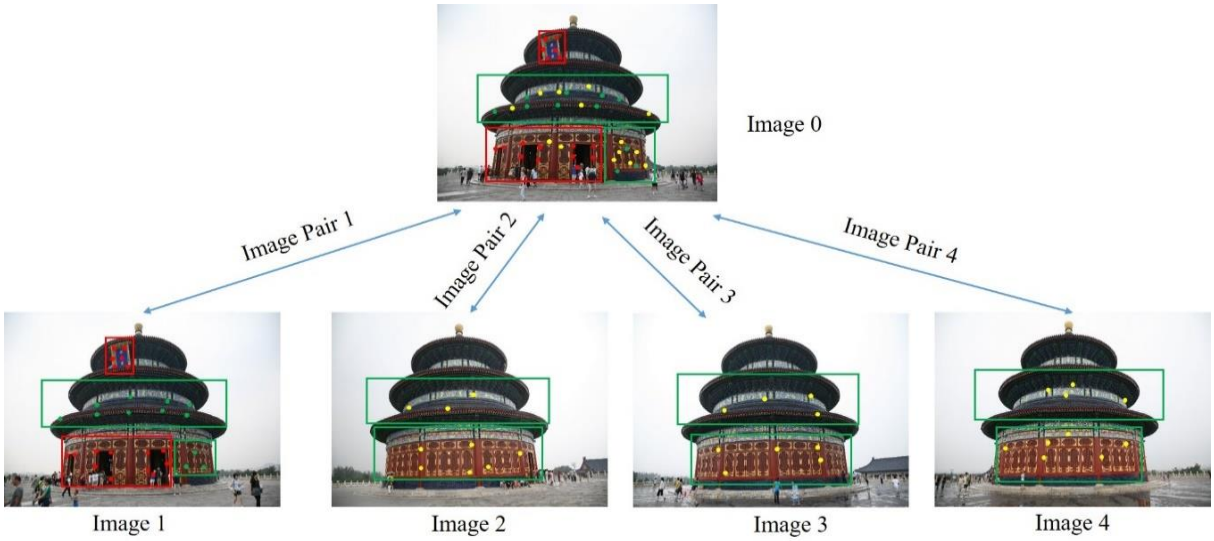


Figure 3.4: Image pairs of non-repetitive and repetitive structure, green boxes denote the RS and red boxed denote the non-RS. For image pair 1, red points are the correspondences from the non-RS, green ones are the correspondences from RS. For image pairs 2,3 and 4, yellow points are the correspondences from RS.

- assuming a constant image size (in pixels), the number of features per image is approximately constant (if the size varies, a normalisation needs to be carried out).

- given a constant overlap between images, overlapping pairs have more conjugate points than non-overlapping pairs after epipolar geometry checking, because the latter do not have any inliers with respect to a central perspective model, as the pair does not overlap.

- a real overlapping image pair has more common image partners with conjugate points, whereas for a non-overlapping pair the further partners of the two images tend to be different.

These hypotheses are used to detect and subsequently eliminate non-overlapping image pairs which passed the 5-point epipolar geometry check.

First, a set of feature points *FP* is constructed: $FP = \{FP_1, FP_2, FP_3, \ldots, FP_N\}$, where *N* is the number of images, FP_i is the set of feature points in the *i*-th image, each represented by an ID (for instance, in Figure 3.4, FP_1 contains the red, green and yellow points). Then, Q_{ij}^i is the set of feature point IDs of the *i*-th image that have matches between the *i*-th and the *j*-th image, such as the red and green points in image pair 1 of Figure 3.4. Now, the difference sets between *FP* and *Q* are constructed by $D_i^j = FP_i \backslash Q_{ij}^i$ for image i and $D_j^i = FP_j \backslash Q_{ji}^j$ for image *j*. Since FP_i is assumed to be approximately constant, and overlapping pairs are assumed to have more matches than non-overlapping ones (see hypotheses above), the number of IDs in both, D_i^j and D_j^i is small for overlapping pairs, and large otherwise. In addition, the IDs in D_i^j with respect to the other images which have correspondences with the *i*-th image is considered as well by generating a vector $\boldsymbol{g}_{ij} = [g_i^1, g_i^2, g_i^3, \ldots, g_i^n]$, where $g_i^j = 0$ and $g_i^k = |\{f \in D_i^j \,\&\&\, f \text{ is a feature matched to the } k\text{-th image}\}|$, $|\,|$ is the operator which returns the number of set elements. Taking Figure 3.4 as an example and studying the vector $\boldsymbol{g}_{ij}$ of image pair 1, the entries of the corresponding vector are

equal to the number of yellow points shown in images 2, 3 and 4. Finally, equation (3.2) computes the degree of repetitive structure RS_{ij} of the i-th and j-th images.

$$RS_{ij} = (|D_i^j| + |D_j^i|)(\boldsymbol{g}_{ij}^{\mathrm{T}}\boldsymbol{g}_{ji})/(|Q_{ij}^i| + |Q_{ji}^j|) \tag{3.2}$$

As mentioned, real overlapping image pairs are assumed to have a small number of elements in the difference set, the value of $\boldsymbol{g}_{ij}^{T}\boldsymbol{g}_{ji}$ should be small as well, and the number of correspondences in the denominator of (3.2) should be large. Thus, the smaller RS_{ij} is, the more probable it is that the image pair does overlap and that the RO is correct, rather than being solely due to repetitive structure.

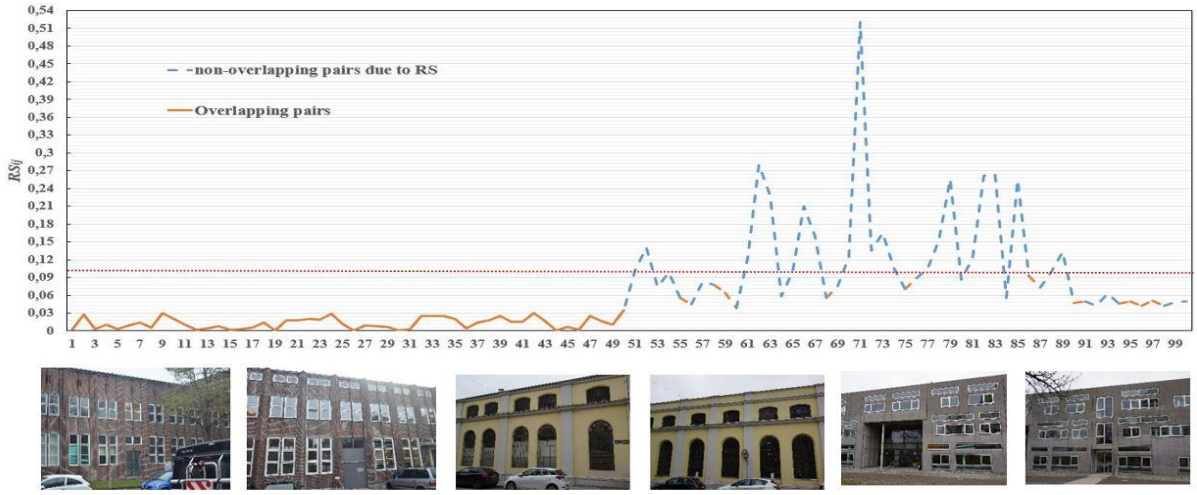


Figure 3.5: Normalized RS_{ij} values of non-RS and RS ROs (upper row) and two corresponding sample images of the three investigated datasets from Wang et al. [2019c] (bottom row). The corresponding ROs are randomly selected from B1, B2 and B3 which contain a lot of repetitive structures. The vertical axis denotes the normalized RS_{ij} values and the horizontal axis is the serial number of ROs that are selected. The solid and dash curve are the non-RS and RS ROs, respectively.

Figure 3.5 illustrates an example for the degree of repetitive structure normalized to the interval [0,1] and shows two sample images of each dataset: some RS and non-RS ROs are randomly selected from the benchmark of Wang et al. [2019c] with lots of repetitive structure and their corresponding normalized RS_{ij} values are shown. The orange parts of the curve indicate overlapping image pairs, i.e. non-RS ROs, and the blue parts represent non-overlapping pairs due to RS. From this figure, it can be found that the ROs with a normalized RS_{ij} value lower than 0.03 are all non-RS ones and ROs whose normalized RS_{ij} value is higher than 0.1 (the red horizontal line) are all RS ROs, whereas the ones between 0.03 and 0.1 stem from either non-RS or RS ROs. Under the assumption that these values hold in general (see section 6 for an experimental investigation), it can be postulated that equation (3.3) can be used to eliminate ROs resulting from repetitive structure:

$$\text{ROs}\begin{cases} non\text{-}RS, \ \ if \ \ nRS_{ij} \leq median(nRS_{ij} \in [0.03, 0.1]) \\ RS, \quad if \ \ nRS_{ij} > median(nRS_{ij} \in [0.03, 0.1]) \end{cases} \tag{3.3}$$

where nRS_{ij} is the normalized RS_{ij} value, and $median(\cdot)$ is an operator to obtain the median value. ROs are non-RS if their corresponding nRS_{ij} is smaller or equal than $median(nRS_{ij} \in [0.03, 0.1])$, and ROs whose nRS_{ij} values are higher than $median(nRS_{ij} \in [0.03, 0.1])$ are considered to be ROs of repetitive structure and are thus eliminated.

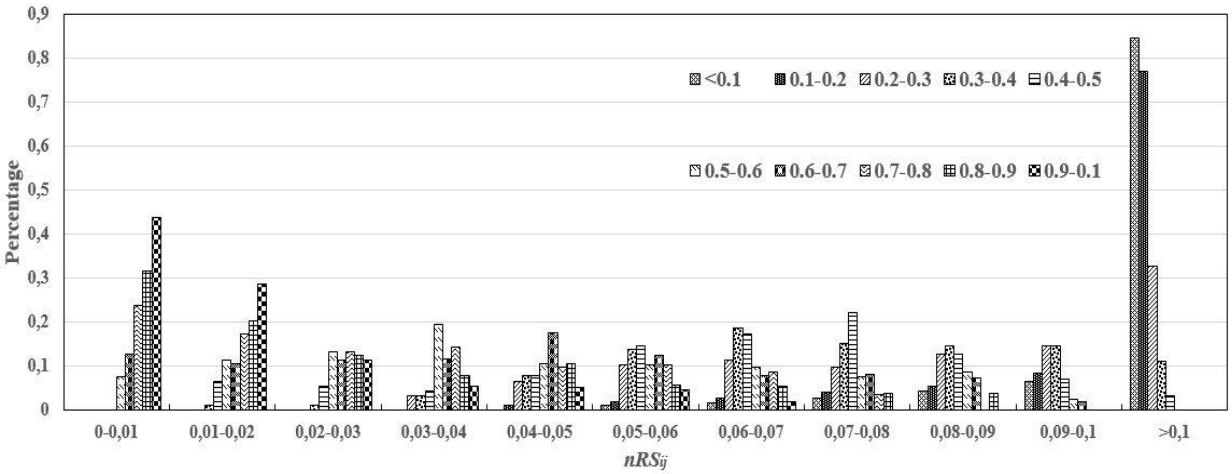


Figure 3.6: Distribution of nRS_{ij} values of image pairs with different overlap ratio.

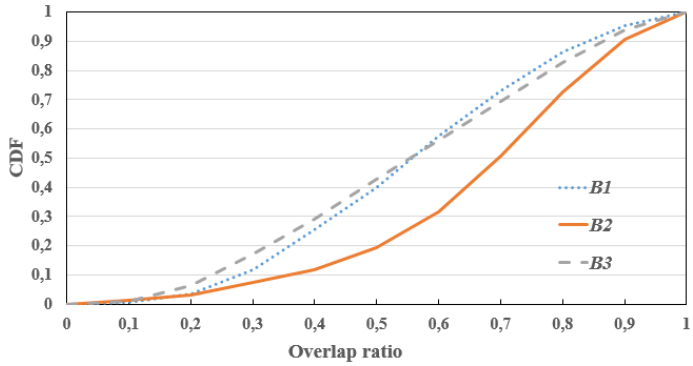


Figure 3.7: The cumulative distribution function (CDF) of overlap ratios from image pairs of corresponding selected ROs. B1, B2 and B3 are the datasets from Wang et al. [2019c], where two sample images of each dataset are shown in Figure 3.5.

The second hypothesis is violated if the image overlap varies; in particular, if the overlapping area is small. In this regard, it can be reasonably argued that in a standard photogrammetric process, ROs of such image pairs are not robust either. Thus, if there are enough images within a block having a proper overlap, it is reasonable to further eliminate pairs with small overlap.

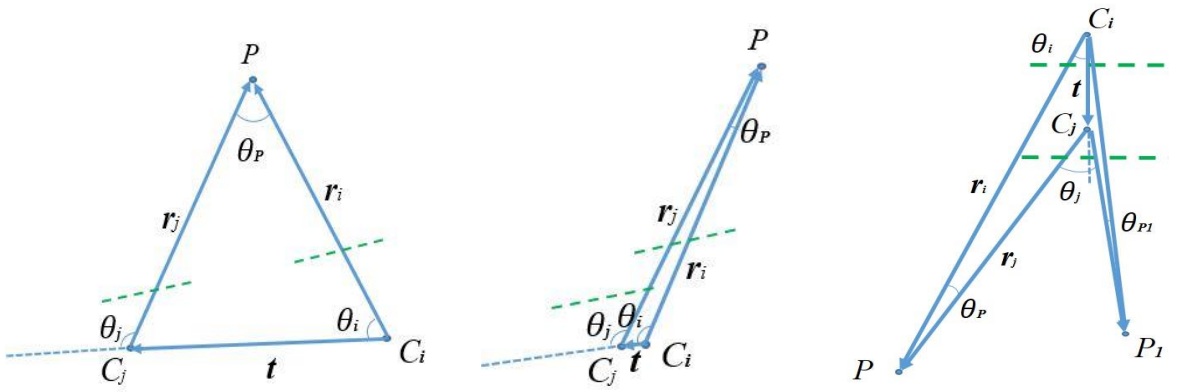
To further investigate this point, images of the benchmark of Wang et al. [2019c] are again used, and the overlap ratio of a certain image pair is estimated by dividing the size of the minimum bounding rectangle containing the correspondences, by the size of the smaller image of this pair. Figure 3.6 shows the distribution of nRS_{ij} values with respect to classes of image pairs categorized by the overlap ratio, where the pattern of a bar denotes an interval of the overlap ratio. From Figure 3.6, image pairs with less than 20 percent of overlap (as bars with overlap ratio values smaller than 0.1 and between 0.1 and 0.2 show) have a nRS_{ij} value higher than 0.05 and for approximately 80% of such image pairs the nRS_{ij} is larger than 0.1. Comparing image pairs of different overlap ratios, it can be seen that the corresponding nRS_{ij} values tend to become smaller as the overlap ratio increases, while most image pairs that overlap more than 30% have a nRS_{ij} value smaller than 0.03.

Assuming the dataset used to be representative for images with repetitive structure, this investigation also indicates that equation (3.3) yields correct results in the large majority of cases.

To demonstrate that most ROs selected according to equation (3.3) do have a reasonable overlap, Figure 3.7 illustrates the cumulative distribution function of overlap ratios for the selected ROs. Only very few image pairs with 10% overlap are selected, and more than 60% of the selected image pairs have an overlap higher than 50%.

3.2.3 Detecting and eliminating RO outliers of very short baselines and baselines parallel to the viewing direction

Critical configurations stemming from very short baselines or baselines parallel to the viewing direction decrease the robustness of global SfM methods in estimating both structure and motion, because the relative translations are no longer estimated with the required precision, which can negatively influence the translation averaging operation [Wang et al., 2019b; Cui and Tan, 2015]. In addition, both cases lead to small intersection angles and thus imprecise coordinates of the ray intersections during triangulation and global translation estimation.



(a) Standard case (b) Very short baseline (c) Baseline parallel to viewing direction

Figure 3.8: Two-view geometry configurations.

Figure 3.8 shows the standard case of two-view geometry with a relatively wide baseline (Figure 3.8a), a VSB case with a very short baseline approximately perpendicular to the viewing direction (Figure 3.8b), and a BPVD case with a long baseline nearly parallel to both viewing directions (Figure 3.8c). Dashed lines denote the image planes, P and P_1 are object points, $\boldsymbol{C_i}$ and $\boldsymbol{C_j}$ are the projection centers of images i and j, $\boldsymbol{t}$ represents the baseline vector from $\boldsymbol{C_i}$ to $\boldsymbol{C_j}$, r_i and r_j are two projection rays, θ_i is the intersection angle of $\boldsymbol{t}$ and r_i, θ_j is the intersection angle of $\boldsymbol{t}$ and r_j, θ_p and θ_{p1} are the intersection angles of corresponding projection rays of object points P and P_1. In the standard case, a reasonable intersection angle θ_p can be obtained and thus the inequality $0 < \theta_i < \theta_j < \pi$ holds, whereas, for the VSB and BPVD, θ_p is very small and the inequality $0 < \theta_i \approx \theta_j < \pi$ can be set up. In addition, in the case of BPVD the smaller the perpendicular distance between an object point and the (extended) baseline $\boldsymbol{t}$ is, the lower is the corresponding intersection angle, e.g., $\theta_{p1} < \theta_p$.

Two equations are presented to distinguish VSB and BPVD cases from standard cases. For a standard case, each pair of conjugate points can contribute as:

$$0 < \theta_i < \theta_j < \pi \Rightarrow cos^{-1}\frac{(\boldsymbol{R_{ij}x_j})^T\boldsymbol{t_{ij}}}{|\boldsymbol{x_j}||\boldsymbol{t_{ij}}|} > cos^{-1}\frac{\boldsymbol{x_i}^T\boldsymbol{t_{ij}}}{|\boldsymbol{x_i}||\boldsymbol{t_{ij}}|} \tag{3.4}$$

$$\text{i.e. } \frac{\boldsymbol{x_i}^T\boldsymbol{t_{ij}}}{|\boldsymbol{x_i}||\boldsymbol{t_{ij}}|} > \frac{(\boldsymbol{R_{ij}x_j})^T\boldsymbol{t_{ij}}}{|\boldsymbol{x_j}||\boldsymbol{t_{ij}}|} \tag{3.5}$$

where, $\boldsymbol{R_{ij}}$ is the relative rotation and $\boldsymbol{t_{ij}}$ is the relative translation (note that in this case the normalized $|\boldsymbol{t_{ij}}|$ equals to 1). $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are the image coordinates of conjugate points as predicted from image matching. These equations can then be simplified as

$$(|\boldsymbol{x_j}|\boldsymbol{x_i}^T - |\boldsymbol{x_i}|(\boldsymbol{R_{ij}x_j})^T)\frac{\boldsymbol{t_{ij}}}{|\boldsymbol{t_{ij}}|} > 0 \tag{3.6}$$

$$cc_{ij}(R) = |\,(\,|\boldsymbol{x_j}|\boldsymbol{x_i}^T - |\boldsymbol{x_i}|(\boldsymbol{R_{ij}x_j})^T)\frac{\boldsymbol{t_{ij}}}{|\boldsymbol{t_{ij}}|}\,| \tag{3.7}$$

Since $0 < \theta_i \approx \theta_j < \pi$ for VSB and BPVD, it can be shown from equation (3.4) that $cos^{-1}\frac{(\boldsymbol{R_{ij}x_j})^T\boldsymbol{t_{ij}}}{|\boldsymbol{x_j}||\boldsymbol{t_{ij}}|} \approx cos^{-1}\frac{\boldsymbol{x_i}^T\boldsymbol{t_{ij}}}{|\boldsymbol{x_i}||\boldsymbol{t_{ij}}|}$, i.e., $(|\boldsymbol{x_j}|\boldsymbol{x_i}^T - |\boldsymbol{x_i}|(\boldsymbol{R_{ij}x_j})^T)\frac{\boldsymbol{t_{ij}}}{|\boldsymbol{t_{ij}}|} \approx 0$. Thus, cc_{ij}(R) should be very close to 0. However, cc_{ij}(R) is far away from 0 when dealing with ROs of normal cases.

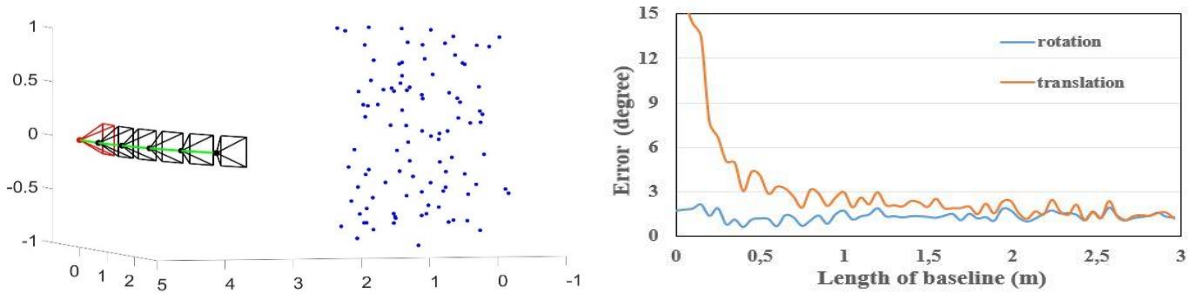
Each pair of correspondences yields one value $cc_{ij}(R)$. To remove the influence of different depths on $cc_{ij}(R)$ (note that object points far away from the projection centers normally yield smaller values for cc_{ij} (R) than those that are closer), this research uses the mean value, $called\ BL_{ij}$, of the top 10% largest $cc_{ij}(R)$ as a criterion to quantify the degree of an image pair having a VSB or BPVD (see Equation (3.8)). The smaller the BL_{ij} is, the higher the probability that the image pair has a very short baseline or a baseline parallel to the viewing direction.

$$BL_{ij} = avg\,(cc_{ij}(R) \in \{\text{top 10\% largest } cc_{ij}(R)\}) \tag{3.8}$$

where *avg*(.) returns the mean value.

As a side note, observe that there exists an implicit assumption that the length of a baseline cannot be equal to zero when decomposing the essential matrix into relative rotation and translations [Longuet-Higgins, 1981; Hartley and Zisserman, 2004]. However, relative rotations can obviously still be computed for image pairs with a zero-length baseline, as this is the task of transforming images into epipolar geometry, and equation (3.8) remains correct in this case; the corresponding derivation can be found in appendix A.

In order to investigate the relevance of this assumption for the presented work, a simulation experiment is designed to investigate whether rotations can still be accurately estimated if the baseline is very short or even has a length of zero.



(a) (b)

Figure 3.9: Simulation experiment. (a) shows the poses of the simulated cameras and the position of object points. The red frame is the fixed camera and the black frames denote the different projection centers of the second camera. (b) shows the error in degree of relative rotation and translation for different baseline lengths.

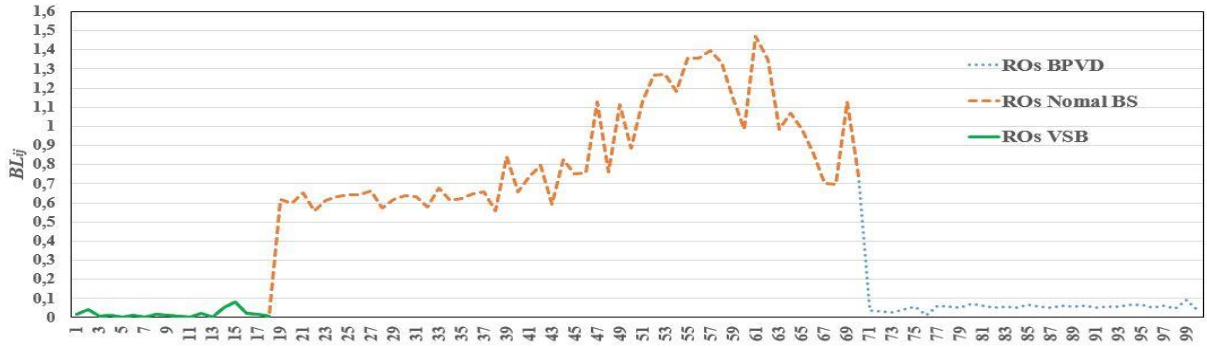
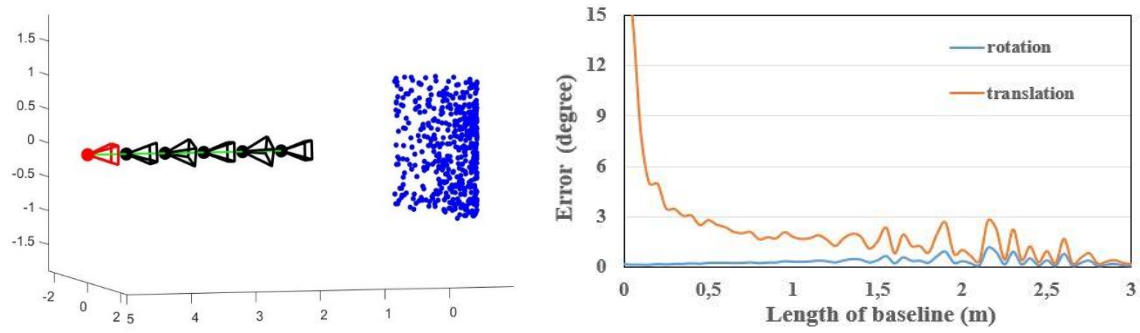


Figure 3.10: BL_{ij} values of ROs from various baselines. The corresponding ROs of very short baselines and normal baselines are randomly select from the tested benchmarks B1, B2 and B3, ROs of BPVD are from UAV1 dataset described in table 5.1 (see Section 5.2). The vertical axis denotes the BL_{ij} values and the horizontal axis shows the serial number of ROs that are selected. The dashed, solid and pointed curves denote normal baseline, very short baseline and baseline parallel to viewing direction, respectively.

As Figure 3.9 (a) shows, a set of 100 3D points is randomly generated in a cube of $[-1,1]^3$. Two cameras are simulated with a focal length of 3500 pixels and an image size of 1200×800 pixels viewing these 3D points. One reference camera is fixed at point (5,0,0) and the second camera moves from this point along an arc (shown by the green line) with its center at (0,0,0) and 5m radius until these two cameras are 3m (arc distance) away from each other. The corresponding rotation matrices are designed by requiring these two cameras to be able to view all 3D points. Based on this setup, image pairs with known exterior orientation parameters are simulated for baselines between 0m and 3m. The image coordinates of the 3D object points are generated via the collinearity equations using Gaussian noise with a standard deviation of 0.2 pixels. The relative orientations of these image pairs are estimated using the 5-point algorithm with the resulting conjugate point coordinates and are compared to the simulated exterior orientation parameters. Since the relative translation is normalized and the scale is unknown, it is only feasible to compare the translation directions. The arc between two cameras is transferred into baseline length. The obtained results are showed in Figure 3.9 (b): The relative rotation error remains stable, while the relative translation error increases as the baseline decreases, which means that the relative rotation can be robustly estimated, while the relative translation cannot, when the baseline is very short.



(a) (b)

Figure 3.11: Simulation experiment along the viewing direction. (a) shows the poses of the simulated cameras and the position of object points. The red frame is the fixed camera and the black frames denote the different projection centers of the second camera. (b) shows the error in degree of relative rotation and translation for different baseline lengths.

Similar to the idea of eliminating RS ROs, random selections of ROs with very short baseline and normal baseline are taken from the benchmark Wang et al. [2019c], as well as ROs with BPVD from the *UAV1* dataset with BPVD ROs illustrated in section 5.2. Equation (3.8) is then employed to compute the corresponding BL_{ij} values. As the curves in Figure 3.10 show, ROs with normal baselines typically generate BL_{ij} values which are much higher than 0.1, whereas the other two cases of baselines have BL_{ij} values that are all below 0.1. Therefore, in this thesis, ROs with BL_{ij} values higher than 0.1 are considered as normal baselines. Thus, there are two remaining possibilities for BL_{ij} values smaller than or equal to 0.1: one is the critical configuration of VSB and the other is ROs of BPVD. As it is known that the RO can be correctly estimated for image pairs of BPVD with a reasonably long baseline, it is advantageous to identify ROs of BPVD in order to keep them as part of the photogrammetric block.

3.2.4 Identifying correct ROs of baselines parallel to the viewing direction

To investigate the precision of BPVD ROs, another simulation is conducted, similar to the one discussed earlier and shown in Figure 3.9. However, this time the simulated camera's motion trajectory is along the camera's viewing direction as the green line in Figure 3.11 (a) shows. Comparing Figure 3.11 (b) with Figure 3.9 (b), significant similarities can be observed, which implies that in the case of baselines parallel to the viewing direction, the relative translation is still imprecise for very small baselines, whereas, it becomes more accurate when the length of the baseline increases. Relative rotations, on the other hand, can be robustly estimated independent of the length of the baseline. Hence, it makes sense that correct BPVD ROs are retained, although they were initially sorted out due to small BL_{ij} values as described in the last subsection.

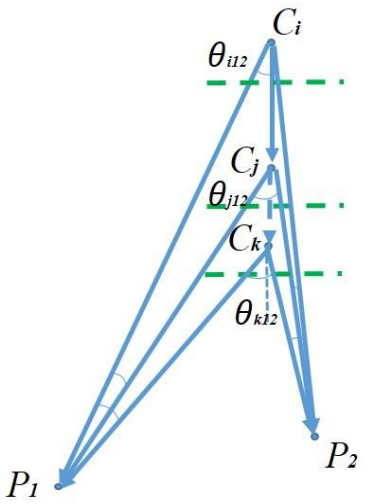


Figure 3.12: Two view geometry constraint for a baseline parallel to the viewing direction.

In Figure 3.12, $\boldsymbol{C_i}$, $\boldsymbol{C_j}$ and $\boldsymbol{C_k}$ are the projection centers of image *i*, *j* and *k*, P_1 and P_2 are object points, θ_{i12}, θ_{j12} and θ_{k12} are the corresponding intersection angle of these object points with respect to the respective projection centers. In the case of BPVD, as Figure 3.12 shows, for any two object points (which are not collinear with any of the projection centers), the intersection angle with the closer projection center should always be larger than that with projection center further away, such that $\theta_{k12} > \theta_{j12} > \theta_{i12}$. Therefore, for every image pair, for example for image *i* and *j*, all such intersection angles for pairs of object points that are projected into different quadrants in the image plane (to avoid intersection angles which are too small) are estimated in the BPVD cases (and in contrast to the VSB cases dealt with above). Subsequently, for each pair of calculated intersection angles θ_{j12} and θ_{i12}, the equation $\theta_{j12} > \theta_{i12}$ can be set up, and the longer the baseline, the larger is the difference between these two angles. The goal is to distinguish short from longer BPVD baselines, as only the longer ones are to be kept. In order to do so, the average $avg\theta$ of all angle differences (θ_{j12}-θ_{i12}) is computed, and pairs with average values larger than 0.1 (in radian) are identified as correct BPVD ROs.

Thus, (3.9) is formulated to conclude the RO selection in terms of critical configurations:

$$\text{ROs}\begin{cases} VSB,\ if\ \ BL_{ij} \le 0.1 \text{ and } avg\theta\ \le 0.1 \\ BPVD,\ if\ \ BL_{ij} \le 0.1 \text{ and } avg\theta\ > 0.1 \\ Normal\ baseline,\quad if\ \ BL_{ij} > 0.1 \end{cases} \tag{3.9}$$

ROs are identified to have a normal baseline if the corresponding BL_{ij} value is larger than 0.1. In contrast, a value smaller than or equal to 0.1 can have two reasons: ROs have a VSB if the corresponding $avg\theta$ value is smaller or equal than 0.1; and ROs are categorized as BPVD cases if $avg\theta$ is larger than 0.1.

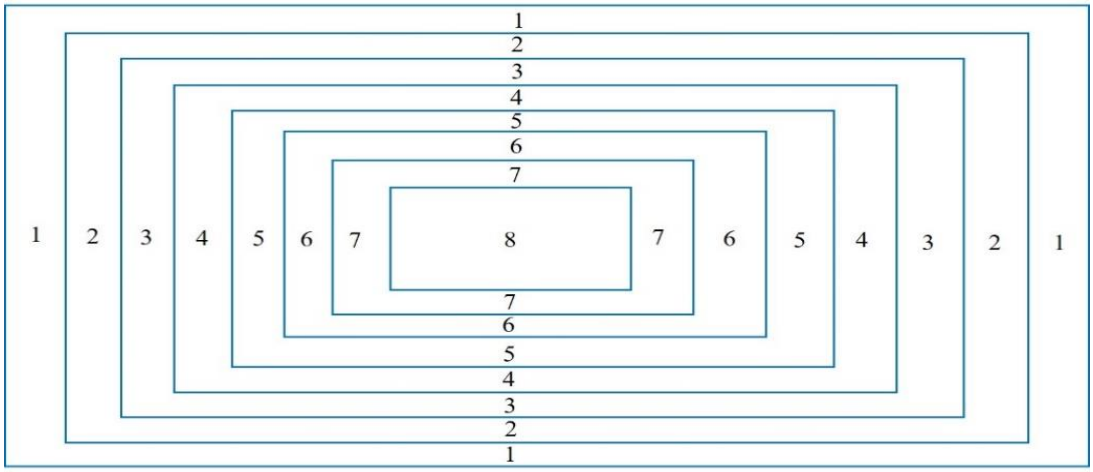


Figure 3.13. Division of images.

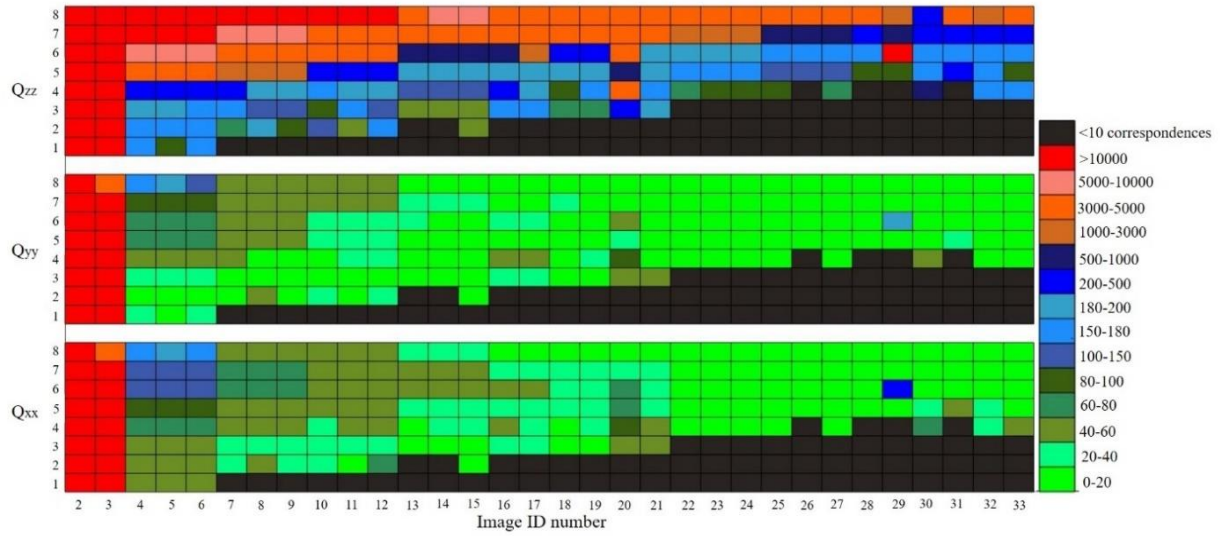


Figure 3.14. Cofactor on X, Y and Z between the first and the rest of the images.

For BPVD cases, another problem exists, even if the baseline is reasonably wide: as the ray intersection angles are small, the depth uncertainty is rather large. As many global SfM pipelines use this depth information to do global translation estimation [Wang et al., 2019a and 2019b; Cui and Tan, 2015; Cui et al, 2015], those translation estimates can become rather uncertain, too. In particular, the global SfM method proposed in this thesis estimates the global unified scale factors for every available image pair by employing the depth values of object points from individual local spatial intersections. Moreover, the refined ROs are fed into the subsequent global SfM pipeline (explained in Chapter 4) for 3D reconstruction in the corresponding experiments. However, Figure 3.12 depicts that BPVD ROs normally lead to small intersection angles for object points which further decrease as the distances between object points and viewing direction becomes smaller. So, in principle, the triangulation results of BPVD image pairs show a higher uncertainty in depth, also referred to as Z direction, than in *X* and *Y* direction (these two directions are defined as within the image space). This effect becomes even more visible, if the distance between object points and

the viewing direction becomes smaller. To demonstrate this behavior, some close range images along viewing direction are taken. In particular, at each station three images are captured by only rotating the camera around the viewing direction, while the principal point is close to the image center (see *CR1* in Table 5.1 and Figure 5.4). The relative orientation between the first image and the rest of the images is estimated (using the 5-point algorithm and RANSAC), and an investigation of the cofactors of *X*, *Y*, *Z* of the triangulating is conducted. In this context, the 2D image planes are evenly divided into eight parts as a rectangular ripple, so that the distances between two neighboring rectangles in vertical and horizontal directions are equal, as Figure 3.13 illustrates.

For the correspondences that are located in the same part of an image, the average values of the cofactors *X*, *Y* and *Z* are computed if more than 10 correspondences exist in the corresponding part. Figure 3.14 visualizes the results, where the horizontal axis shows the image ID and the vertical axis denotes the cofactors of *X*, *Y* and *Z*, namely, *Qxx*, *Qyy* and *Qzz* for each part of the image. As it can be seen from this result, images 2 and 3 have quite large cofactors on these three directions. Because the images 1, 2 and 3 are taken from the same photogrammetric station by only rotating the camera, the baselines between these images are very short (almost equal to 0), resulting in relative translations that are totally wrong. As the camera moves away along the viewing direction, all the cofactors tend to decrease. As what one can expect: first, *Qzz* is always larger than *Qxx* and *Qyy*; second, the points that are close to the image center show a much worse *Qzz*. Although *Qzz* is never as good as *Qxx* and *Qyy* in BPVD image pairs, one could select the most reliable object points from these "not good" ones for global SfM. In this thesis, only the correspondences such that the corresponding $Qzz < 10 \times max\,(Qxx, Qyy)$ ($max\,(\cdot)$ returns the maximum item) are kept.

3.3 Discussion

This chapter addresses time efficient and robust methods for pre-processing steps, whose results are considered as input for the subsequent global image orientation. Based on a random k-d forest, first a time efficient strategy is introduced to determine overlapping image pairs. The goal is to only match and compute the relative orientations for those pairs which actually show a sufficient overlap. Then, novel methods are presented to eliminate blunders in ROs for conducting robust global image orientation. In particular, the compatibility of triplets regarding relative rotation and translation errors is checked, and RO outliers due to RS (repetitive structure) and VSB (very short baseline) are investigated. Criteria for these two cases are introduced, and as the latter criterion is sensitive to BPVD (baseline parallel to the viewing direction) pairs, also a new criterion for this case is proposed accordingly. Nevertheless, the following challenges are still open, posing potential limitations, and need to be overcome to improve the robustness:

1) While enough points are currently guaranteed to be available for the computation of the relative orientation parameters, both the point distribution in the overlapping area and error propagation are not taken into account to detect potential numerical problems in parameter

estimation. Neither is the investigation of a homography $\boldsymbol{H}$ as an alternative to the fundamental matrix $\boldsymbol{F}$ performed to overcome difficulties stemming from planar objects. Another possible option is to use a trifocal tensor instead of the essential or fundamental matrix to tie together three images rather than two as basic building blocks. While this solution is more reliable by considering more geometrical constraints, the corresponding complexity must be considered, because the number of triplets can theoretically be cubic in terms of the number of images.

2) The solutions presented in this section contain a number of assumptions and some free parameters which need to be determined in advance. These free parameters are selected empirically, and the most reasonable ones are generalized to all datasets. This strategy works very well and is thoroughly demonstrated on various datasets in the experimental section. However, in real scenarios, the obtained dataset can be more complex than those used in this work, which may require that some of these free parameters have to be re-adjusted.

4 Global image orientation

The Preprocessing outlined above aims at efficiently providing inputs that is largely free of blunders for the subsequent image orientation task. In this chapter, these inputs are represented by means of a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, known as viewgraph, in which each vertex in $\mathcal{V}$ indicates an image and an edge (i, j) in $\mathcal{E}$ denotes the relative orientation of images i and j which was estimated and successfully passed the previous RO outlier elimination process. Furthermore, the inputs also contain the correspondences (x_i, x_j) of each image pair (i, j) and the 3D coordinates of the tie points generated by these correspondences. To obtain an image orientation solution in a fast and robust way, this chapter is devoted to a detailed explanation of newly developed global image orientation methods using these inputs.

4.1 General Overview

This section gives a brief overview of the proposed global image orientation solution. Analogous to the majority of conventional global strategies, a two-step strategy of first estimating global rotations and subsequently solving global translations is applied in this thesis. As shown in Figure 1.2, after the introduction of global rotation estimation (4.2), a novel global translation method is presented (4.3). Then, the refinement using bundle adjustment is described in Section 4.4. Finally, a related discussion is given in Section 4.5.

In the first part (4.2), the used global rotation estimation method is discussed. The corresponding rotation matrix preliminaries are given, before the robust global rotation estimation algorithm is presented. While the main contributions of this thesis do not cover the issue of global rotation estimation, for the completeness of the whole method, the popular global rotation estimation method of Chatterjee and Govindu [2013] was selected and is explained here. Two reasons are given for this choice: *first*, to make a fair comparison of the relevant contributions on global translation estimation, it is important that the global rotation results are computed with the same global rotation estimation method; *second*, their work is widely used in many state-of-the-art global image orientation algorithms and considered to be capable of providing reliable rotations for large numbers of images.

In the following two parts, given the already estimated global rotations, a newly developed global translation estimation method which utilizes relative translations together with tie points is introduced.

The fourth part comprises some details of the final robust bundle adjustment (4.4) and the last part concludes the presented image orientation methods and discusses the corresponding limitations and possible solutions.

4.2 Global rotation estimation

This section focuses on determining global rotations from pairwise relative rotations. To start with, in section 4.2.1, some fundamental preliminaries of rotations are reviewed (e.g., characteristics of the relevant Lie group and Lie algebra), before the problem of solving global rotations is stated. Section 4.2.2 is dedicated to the detailed procedure for global rotation estimation using the method of Chatterjee and Govindu [2013], in which a hybrid coarse-to-refined strategy using the L_1 norm and iterative reweighted least squares is adopted. Finally, while this thesis does not intend to contribute to the task of global rotation estimation, some practical difficulties are studied and corresponding hints for solving them are discussed in 4.2.3.

4.2.1 Rotation preliminaries and problem statement

All 3×3 rotation matrices $\boldsymbol{R}$ form a closed group known as the Special Orthogonal group $SO(3)$, which is a differentiable Riemannian manifold, i.e., $SO(3)$ is a Lie group which in turn is the basis for efficient methods for rotation estimation. Apart from the standard characteristics of a group, a Lie group has a smooth differentiable structure providing the additional advantage that the product and inverse operations are differentiable mappings. The local neighbor of a point in a Lie group is topologically equivalent to a vector space, i.e., a Lie algebra $\mathfrak{so}(3)$. The mapping techniques of rotation matrices between $SO(3)$ and $\mathfrak{so}(3)$ are denoted as logarithm and exponential mapping. The projection of $SO(3)$ to $\mathfrak{so}(3)$ is given by:

$$log(\boldsymbol{R})=[\boldsymbol{\varrho}]_\times \text{ , } \boldsymbol{\varrho} = \arcsin(\|\boldsymbol{w}\|_2)\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}, \boldsymbol{w} = \frac{\boldsymbol{R}-\boldsymbol{R}^T}{\boldsymbol{2}} \tag{4.1}$$

in which $\boldsymbol{\varrho}$ is a 3-dimensional vector and $\mathfrak{so}(3)$ is the relevant skew-symmetric matrix indicated as $[\boldsymbol{\varrho}]_\times$

$$[\boldsymbol{\varrho}]_\times = \begin{pmatrix} 0 & -\varrho 3 & \varrho 2 \\ \varrho 3 & 0 & -\varrho 1 \\ -\varrho 2 & \varrho 1 & 0 \end{pmatrix}$$

The corresponding back projection which maps $\mathfrak{so}(3)$ into $SO(3)$ is given the by exponential operation:

$$\boldsymbol{R}= exp([\tilde{\boldsymbol{\varrho}}]_\times)= \boldsymbol{I} +\sin(\alpha)\, [\tilde{\boldsymbol{\varrho}}]_\times+(1-\cos(\alpha))\, [\tilde{\boldsymbol{\varrho}}]_\times^2 \tag{4.2}$$

where $\boldsymbol{I}$ is an identity matrix, $(\boldsymbol{\varrho}, \alpha)$ is in axis-angle representation of $\boldsymbol{R}$ with $\boldsymbol{\varrho} = \alpha\tilde{\boldsymbol{\varrho}}$ and $\|\tilde{\boldsymbol{\varrho}}\|_2=1$.

The inherent bi-invariant distance between two rotation matrices on $SO(3)$ can be formulized as $d(\boldsymbol{R_i}, \boldsymbol{R_j}) = \|\log(\boldsymbol{R_i}\boldsymbol{R_j^{-1}})\|_F = \|\log(\boldsymbol{R_j}\boldsymbol{R_i^{-1}})\|_F$, where $\|.\|_F$ is the *Frobenius* norm. The global

rotation estimation problem can then be stated as following: Given a set of relative rotations $\{\boldsymbol{R_{ij}} \mid (i, j) \in \mathcal{E}\}$, the goal is to determine the global rotations $\boldsymbol{R}_{global} = \{\boldsymbol{R_1}, \ldots, \boldsymbol{R_N}\}$ with respect to a given frame of reference such that

$$\underset{\boldsymbol{R}_{global} \in SO(3)}{\operatorname{argmin}} \sum_{(i,j) \in \mathcal{E}} d^2(\boldsymbol{R_{ij}}, \boldsymbol{R_j R_i^{-1}}) \tag{4.3}$$

Considering just one relative rotation $\boldsymbol{R_{ij}} = \boldsymbol{R_j R_i^{-1}}$ indicated by the edge $(i, j) \in \mathcal{E}$, according to the Baker-Campbell-Hausdorff-formula [Gilmore, 1974], the first–order approximation of the corresponding Lie algebraic relation can be written as $\boldsymbol{\varrho_{ij}} = \boldsymbol{\varrho_j} - \boldsymbol{\varrho_i}$. The global rotations are further denoted by angle-axis representation as $\boldsymbol{\varrho_{global}} = \{\boldsymbol{\varrho_1}, \ldots, \boldsymbol{\varrho_N}\}$. In consequence, for the given edge $(i, j) \in \mathcal{E}$, the relationship is formulated as

$$\boldsymbol{\varrho_{ij}} = \boldsymbol{\varrho_j} - \boldsymbol{\varrho_i} = \underbrace{[\cdots -\boldsymbol{I}_{3\times3} \cdots \boldsymbol{I}_{3\times3} \cdots]}_{\boldsymbol{Av}_{ij}} \boldsymbol{\varrho_{global}} \tag{4.4}$$

where in $\boldsymbol{Av}_{ij}$, $\boldsymbol{I}$ and $-\boldsymbol{I}$ are set as 3×3 blocks in the corresponding location of j and i respectively.

$$\boldsymbol{\varrho_{rel}} = \boldsymbol{A_v \varrho_{global}} \tag{4.5}$$

By constructing all edges in $\mathcal{E}$ in the form of (4.4), equation (4.5) is obtained, where $\boldsymbol{\varrho_{rel}}$ is the vector obtained from stacking all relative rotations $\boldsymbol{\varrho_{ij}}$ and $\boldsymbol{A_v}$ is the coefficient matrix obtained from stacking all the related matrices $\boldsymbol{Av}_{ij}$.

To solve equation (4.5), the discrepancy between observations $\boldsymbol{R_{ij}}$ and the current estimation is optimized in the Lie algebra for all relative rotations. Based on a *Gauss-Markov model*, equation (4.5) can be written in the form of $\boldsymbol{e_v} = \boldsymbol{A_v}\Delta\boldsymbol{\varrho_{global}} - \Delta\boldsymbol{\varrho_{rel}}$, where $\Delta\boldsymbol{\varrho_{rel}}$ is the corresponding collection of $\Delta\boldsymbol{\varrho_{ij}}$, following the estimation of step 3 as algorithm 4.1 shows, the individual rotations are updated by exponentially mapping the Lie algebraic update item $\Delta\boldsymbol{\varrho_i}$ to the Lie group $SO(3)$. This procedure guarantees that the algorithm always provides a solution which is located on the rotation manifold.

Algorithm 4.1 Lie-Algebraic global rotation estimation

Input: $\{\boldsymbol{R_{ij}} \mid (i, j) \in \mathcal{E}\}$

Output: $\boldsymbol{R}_{global} = \{\boldsymbol{R_1}, \ldots, \boldsymbol{R_N}\}$ with respect to a given frame of reference

Initialization: Obtain an initial value for $\boldsymbol{R}_{global}$

While $\|\Delta\boldsymbol{\varrho_{rel}}\| > \epsilon$ **do**

1. $\Delta\boldsymbol{R_{ij}} = \boldsymbol{R_j^{-1} R_{ij} R_i}$
2. $\Delta\boldsymbol{\varrho_{ij}} = log(\Delta\boldsymbol{R_{ij}})$
3. Solve $\boldsymbol{A_v}\Delta\boldsymbol{\varrho_{global}} = \Delta\boldsymbol{\varrho_{rel}}$
4. $\boldsymbol{R_i} = \boldsymbol{R_i} exp(\Delta\boldsymbol{\varrho_i})$, $i \in [1, N]$

End

4.2.2 Robust solution of global rotations

To achieve a robust solution, Chatterjee and Govindu [2013] proposed a coarse to refined strategy. In particular, the initial solution is obtained using the robust L_1 norm, and then the solution is further improved by an iteratively reweighted least squares approach using the Huber-like estimator.

Robust coarse solution. As the Lie algebra is a vector space, the problem of robust optimization in the Lie algebra is analogous to the robust estimation of a linear equation system. Inspecting the conventional linear problem of $Ax = b$ where $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ (m>n), the unknown x can be determined if A is of full rank. However, the difficulty of solving it varies when the observations are corrupted by noise and outliers, resulting in $b = Ax + e$, where e is the residual indicating the difference between the model and the given observations. Step 3 $\boldsymbol{A_v}\Delta\boldsymbol{\varrho_{global}} = \Delta\boldsymbol{\varrho_{rel}}$ in algorithm 4.1 can be solved by

$$\underset{\Delta\boldsymbol{\varrho_{global}}}{\operatorname{argmin}} \left\| \boldsymbol{e_\varrho} \right\|_{L_1} \quad \Leftrightarrow \quad \underset{\Delta\boldsymbol{\varrho_{global}}}{\operatorname{argmin}} \left\| \boldsymbol{A_v}\Delta\boldsymbol{\varrho_{global}} - \Delta\boldsymbol{\varrho_{rel}} \right\|_{L_1} \tag{4.6}$$

In equation (4.6), $\boldsymbol{e_\varrho} = \boldsymbol{A_v}\Delta\boldsymbol{\varrho_{global}} - \Delta\boldsymbol{\varrho_{rel}}$, $\Delta\boldsymbol{\varrho_{global}} \in \mathbb{R}^{3|\mathcal{V}|}$ and $\Delta\boldsymbol{\varrho_{rel}} \in \mathbb{R}^{3|\mathcal{E}|}$, since the number of unknown global rotations is equal to the number of nodes in $\mathcal{G}$, i.e., $|\mathcal{V}|$, the number of known observations is equal to the number of edges in $\mathcal{G}$, i.e., $|\mathcal{E}|$, and each $\Delta\boldsymbol{\varrho_i}$ is a 3-dimensional vector. Evidently, the number of edges $|\mathcal{E}|$ is much larger than the number of vertices $|\mathcal{V}|$. The relevant L_1 norm embedded in algorithm 4.1 is denoted as *L1RA,* for which each row of $\boldsymbol{A_v}$ contains only two non-zero entries {-1, +1}, characterizing $\boldsymbol{A_v}$ as extremely sparse which allows to compute the solution efficiently. In addition, this L_1 norm problem is solved with the alternating direction method of multipliers (ADMM) as a least unsquared deviations minimizer, more details of implementing ADMM for L_1 minimization can be found in Boyd et al. [2010].

Refined solution. While the L_1 minimization mentioned above offers a robust rotation estimation in the presence of outliers, the estimation result can be further improved by considering the problem of robust global rotation estimation as M-estimator modifications of least squares estimation. However, if the solution obtained by L_1 norm is approximately accurate, the residuals for individual relative rotations provide a good indication of the reliability of the input $\{\boldsymbol{R_{ij}} \mid (i, j) \in \mathcal{E}\}$. This information is utilised to iteratively solve a robust weighted least squares problem that minimizes the discrepancy between relative rotations and the estimated global rotations.

Investigating the linear equation system $Ax = b$, one widely known solution is the standard least squares loss function $e^T e$ where $e = Ax - b$. In the sense of coping with outliers, a robust version of this loss function is used, i.e., $\sum_i \rho(||e_i||)$, where e_i is the i-th entry of the error vector e and $\rho(.)$ is the robust Huber-like loss function $\rho(x) = {x^2}/{(x^2 + c^2)}$. To obtain a value for x which minimizes $\sum_i \rho(||e_i||)$, the following is derived:

$$\underset{x}{\operatorname{argmin}} \sum_i \rho(||e_i||) = \underset{x}{\operatorname{argmin}} \sum_i {e_i}^2 \Big/ ({e_i}^2 + c^2)$$

$$\Rightarrow \frac{\partial \sum_i \rho(||e_i||)}{\partial x} = \frac{\partial \sum_i \rho(||e_i||)}{\partial e} \frac{\partial e}{\partial x} = 0$$

$$\Rightarrow A^T \phi(e) Ax = A^T \phi(e) b \tag{4.7}$$

in which $\phi(e)$ is a diagonal matrix with $\phi(i,i) = {c^2} \Big/ ({e_i}^2 + c^2)^2$. The system of equations (4.7) is a non-linear optimization due to the dependency of $\phi(e)$ on x through e. Nevertheless, such problems are typically addressed with an iterative scheme. Assuming x is fixed, the vector e can be obtained with $e = Ax - b$ and $\phi(e)$ is then fixed. Subsequently, the relevant minimization problem becomes $\underset{x}{\operatorname{argmin}} (Ax - b)^T \phi(e)(Ax - b)$, and its solution is denoted as $x = (A^T \phi(e) A)^{-1} A^T \phi(e) b$. Given this x, $\phi(e)$ can in turn be re-estimated. So, the alternation between computing x (fixing $\phi(e)$) and $\phi(e)$ (fixing x) is recursively conducted until it converged. This strategy is popularly known in literature as *Iteratively Reweighted Least Squares* (*IRLS*) and is intuitively stated as Algorithm 4.2.

Algorithm 4.2 Iteratively Reweighted Least Squares (*IRLS*)

Set an initial value to x

While $||x - x_{prev}|| > \epsilon$ **do**

1. $x_{prev} = x$
2. $e = Ax - b$
3. $\phi = \phi(e)$
4. $x = (A^T \phi A)^{-1} A^T \phi b$

End

It is worth to note that the *IRLS* method can also be applied to robustly solve the Lie algebraic linear problem of step 3 of algorithm 4.1. Although *IRLS* can provide a good solution, it essentially is a greedy algorithm and requires a good initialization of x. In absence of a good initial guess, the intermediate reweighting $\phi(e)$ may not be informative enough and the method may not converge to a reliable final estimation. Since the *L1RA* method shows a high efficacy and offers a good estimate of $\boldsymbol{R}_{global}$, the results of this method are utilized as the initialization for a subsequent robust global rotation estimation using the *IRLS* algorithm.

Algorithm 4.3 Robust Global Rotation Estimation (*L1RA-IRLS*)

L1RA step:

- Compute the initialization of $\boldsymbol{R}_{global}$
- Run Algorithm 4.1 by solving step 3 using Equation (4.6)

IRLS step:

- Set the initial guess of $\boldsymbol{R}_{global}$ as the output of *L1RA*
- Run Algorithm 4.1 by solving step 3 using Algorithm 4.2

The complete robust global rotation estimation of Chatterjee and Govindu [2013] is stated by Algorithm 4.3 (denoted as *L1RA-IRLS*). Both components are demonstrated to be crucial for the estimation process. While the *L1RA* is devoted to offer a good initialization as it is an efficient method for robust estimation, the *IRLS* step is necessary to appropriately weight the uncertainty information from individual relative rotation observations to provide an accurate solution.

4.2.3 Discussion

According to the presented details, the employed global rotation averaging method has several practical characteristics: *First*, the initialization of step *L1RA*. Chatterjee and Govindu [2013] compute the initial guess by using a randomly selected spanning tree. *Second*, the iteration convergence criterion of both steps (which is set to $\epsilon = 10^{-3}$). In the hybrid scheme of *L1RA-IRLS*, to provide an initialization for the *IRLS* step, this convergence criterion is not applied in the *L1RA* stage. In contrast, running *L1RA* for 5 iterations is sufficient to achieve an estimation of $\boldsymbol{R}_{global}$ which fulfills the requirements of the following *IRLS* step (the corresponding tuning parameter c is selected as 5°).

Despite the popularity of the described method, two difficulties still should be taken care of: *first*, the gauge ambiguity. The rotation of the first image (in the image set) is set to be a 3×3 identity matrix to remove the gauge ambiguity of the reference system. This is in fact not a good choice if the first image has a week connection to the photogrammetric block. *Second*, robustness. It is actually of special notice that although the robustness of the method presented by Chatterjee and Govindu [2013] has already been widely demonstrated, their method can further benefit significantly if observations with less outliers are provided (e.g., the presented RO robustification procedure outlined in Chapter 3), because the initialization of *L1RA* has a significant influence on the accuracy that can easily be corrupted by blunders within the set of relative rotations. In addition, a widely used idea to obtain a reliable solution is recommended here, namely, outliers should be detected in the *IRLS* step and then a standard least squares adjustment with all the inliers having equal weights is carried out.

4.3 Global translation estimation

While the previous section focused on the estimation of global rotations, this section concentrates on the other type of exterior orientation parameters: global translations. In particular, a new global translation estimation method using tie points within triplets and relative translations is investigated.

4.3.1 Problem statements and relevant function model

For each image i in the viewgraph, this section aims to estimate its projection center $\boldsymbol{C}_i$ in a consistent global coordinate system (e.g. in the coordinate system of the first image). Each edge

(i, j) of the viewgraph encodes the relative rotation $\boldsymbol{R_{ij}}$ and relative translation $\boldsymbol{t_{ij}}$ of the two images i and j; $\| \boldsymbol{t_{ij}} \| = 1$. These parameters are constrained by equation (4.8):

$$\begin{aligned} \boldsymbol{R_{ij}} &= \boldsymbol{R_j R_i^{-1}} \\ \lambda_{ij}\, \boldsymbol{t_{ij}} &= \boldsymbol{R_i}(\boldsymbol{C_j} - \boldsymbol{C_i}) \end{aligned} \tag{4.8}$$

Referring to equation (4.8), at this stage, $\boldsymbol{R_{ij}}$, $\boldsymbol{t_{ij}}$ and $\boldsymbol{R_i}$ are already known from relative orientation and global rotation estimation (Chapter 4.2), respectively: $\boldsymbol{t_{ij}}$ is the relative translation vector pointing from the projection center of image i to the one of image j (defined in the local coordinate system of *i-th* image), and $\boldsymbol{R_i}$ is the global rotation matrix of image i. The remaining unknowns are the scale factors λ_{ij}, which must be uniquely determined up to a global gauge ambiguity and the coordinates of the projection centers $\boldsymbol{C_i}$ (i, j =1, 2, 3…., N, where N is the number of images in the viewgraph). Collecting all the edges $\mathcal{E}$ in the viewgraph $\mathcal{G}$ using the relationship of equation (4.8), one can set up a linear equation system for global translation estimation. The challenge then is to estimate the global scale factors λ_{ij} for every edge in $\mathcal{G}$. As Figure 4.1 implies, for one image pair, the values of the 3D tie point coordinates are proportional to the length of the baseline. For each image, all the overlapping images are considered first, before the corresponding 3D tie point coordinates are investigated to determine consistent scale factors for the related image tuple. As many tuples are generated as there are images; this step is carried out for each image separately. Subsequently, the different tuples are connected and global consistent scales are derived before determining the projection center coordinates $\boldsymbol{C_i}$.

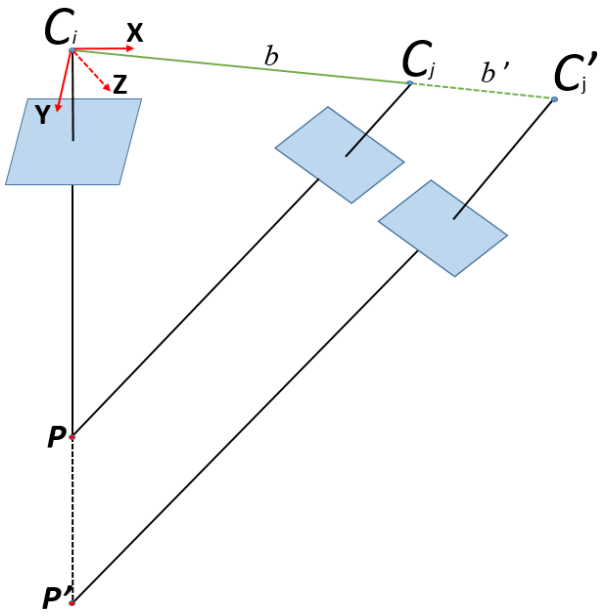


*Figure 4.1: Local pairwise triangulation. $\boldsymbol{C_j}$: projection center of image j if the length of the baseline is defined as b. **P**: the position of a tie point under these circumstances. $\boldsymbol{C_j}$', **P**': Projection center of image j and tie point if the length of the baseline is changed to be b'.*

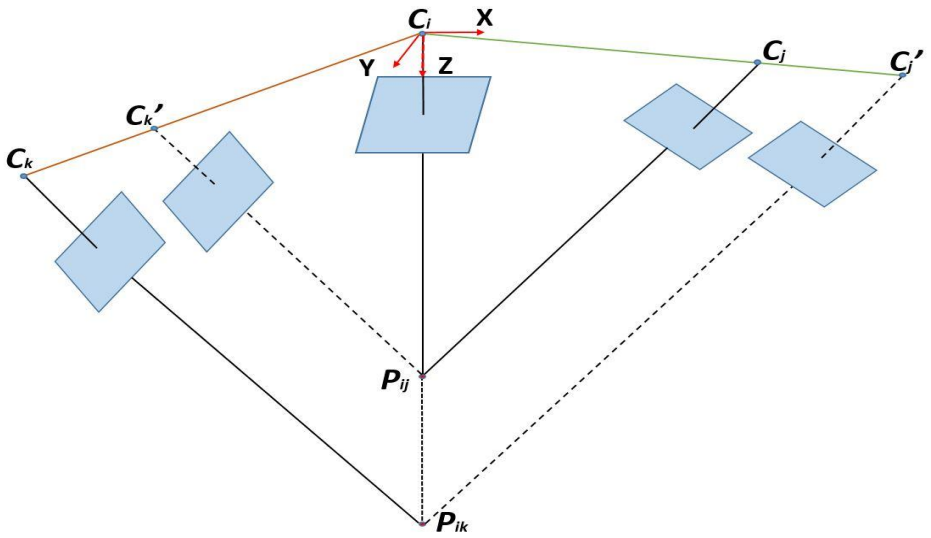


Figure 4.2: Global scale factor ambiguity.

4.3.2 Determination of globally consistent scale factors

Consistent scale factors per image tuple

With reference to Figure 4.1 let the origin of the coordinate system be the projection center of image *i* and its rotation matrix to be the identity matrix. Assuming that the (unknown) baseline length corresponds to $\|\boldsymbol{C}_i\,\boldsymbol{C}_j\|$, the coordinates of object point $\boldsymbol{P}$ can be computed from the image coordinates of its corresponding homologous points. Obviously, if the projection center of image *j* is extended to C_j', object point $\boldsymbol{P}$ will move to $\boldsymbol{P}'$ (Figure 4.1) and it is easy to establish equation (4.9) using ||.|| to denote the length of the respective vectors. As the object point should always be located in front of the camera, the value of the object point's *Z*-coordinate is certainly larger than 0 and the ratio η_{ij}^{i} of the two Z values (Z_p and $Z_{p'}$) can be used instead of $\|\,\boldsymbol{C}_i\,\boldsymbol{P}\,\|/\|\,\boldsymbol{C}_i\,\boldsymbol{P}'\,\|$. This makes the proposed method more computationally efficient, because it is not necessary to calculate the more complex Euclidean distance between each object point and the reference projection center.

$$\frac{\|\boldsymbol{C_i}\,\boldsymbol{C_j}\|}{\|\boldsymbol{C_i}\,\boldsymbol{C_j'}\|} = \frac{\|\boldsymbol{C_i}\,\boldsymbol{P}\|}{\|\boldsymbol{C_i}\,\boldsymbol{P'}\|} = \frac{Z_p}{Z_{p'}} = \eta_{ij}^{i} \tag{4.9}$$

Figure 4.2 offers a direct illustration of the ambiguity resulting from inconsistent scale factors and its solution. Assuming an object point to be visible in three images (with projection centers $\boldsymbol{C}_i$, $\boldsymbol{C}_j$, $\boldsymbol{C}_k$), it can be used for scale transfer from image pair (*i*, *j*) to image pair (*i*, *k*). The corresponding ray intersections must result in identical coordinates for the same tie points in object space, which can be achieved by selecting a proper scale factor for the image pair (*i*, *k*).

Now, by considering all image pairs in the viewgraph which contain image *i* (denoted as reference image), the solution for a consistent scale factor η_{ij}^{i} of each image pair (*i*, *j*) in the local coordinate system of image *i* is presented; For each object point $\boldsymbol{P}$ that can be viewed on both image pairs (*i*, *j*) and (*i*, *k*), equation (4.10) is set up:

$$\frac{\eta_{ij}^{i}}{\eta_{ik}^{i}} = \frac{Z_{ik}}{Z_{ij}} = r_{jk}^{i} \tag{4.10}$$

where r_{jk}^{i} is constant with respect to these three images.

First of all, each triplet is inspected independently. For each three-ray point, a value r_{jk}^{i} can be calculated. To achieve a reliable solution, first the mean value and standard deviation of all r_{jk}^{i} values is calculated. Then, the points with r_{jk}^{i} values within the interval of twice the standard deviation are kept, and the average $\bar{r}_{jk}^{i}$ is computed from the remaining r_{jk}^{i} values. If the resulting number of inliers r_{jk}^{i} is below a threshold *pt*, this triplet is not considered any further.

Next, all images connected to the reference image *i* are considered which allows to solve all the corresponding scales simultaneously. However, equation (4.10) is in the form of a non-linear constraint. Therefore, the logarithm of both sides is taken to formulate it as a linear equation:

$$log\eta_{ij}^{i} - log\eta_{ik}^{i} = log\bar{r}_{ij}^{i} \tag{4.11}$$

By setting up equations of type (4.11) for all image pairs (*j*, *k*) which overlap with image *i*, the linear equation system (4.12) is built, where $\boldsymbol{X}_Z$ and $\boldsymbol{b}_Z$ are vectors. $\boldsymbol{X}_Z$ contains the $\log\eta_{ij}^{i}$ (considered as unknowns) and $\boldsymbol{b}_Z$ the $\log\bar{r}_{jk}^{i}$ values (computed as described above); $\boldsymbol{A}_Z$ is a sparse matrix with rank deficiency 1, containing two non-zero elements per row, 1 and -1. The rank deficiency expresses the fact that the scale can be determined only up to a common value. To remove this ambiguity, for the image pair with the largest number of correspondences $\log\eta_{ij}^{i} = 1$ is used as a constraint.

$$\boldsymbol{A}_Z\boldsymbol{X}_Z = \boldsymbol{b}_Z \tag{4.12}$$

$$arg\ \min_{X_Z} \|\boldsymbol{A}_Z\boldsymbol{X}_Z - \boldsymbol{b}_Z\|_2 \tag{4.13}$$

To obtain the optimal solution for the overdetermined system in equation (4.12), the standard least-squares estimation is adopted according to equation (4.13). For each reference image, it is required to solve the corresponding linear equation system (4.12). $\boldsymbol{A}_Z$ is relatively small when solving for each image independently and the solution can be estimated efficiently. Alternatively, a large equation system could be also built by taking into consideration all images simultaneously. However, $\boldsymbol{A}_Z$ would become very large and it would be inefficient to solve equation (4.13). In this thesis, solving equation (4.12) for each reference image independently is advocated. Furthermore, it is easy to parallelize this procedure, which means several individual image can be solved simultaneously.

After determining consistent scale factors within tuples, translation outlier detection, described in section 3.2.1, is conducted to obtain more robust global scale factors and translations.

Consistent scale factors across image tuples

By applying the above approach, not too much effort is required to obtain consistent scale factors for each image tuple with one image as reference image. However, these scale factors are not consistent across tuples. It is thus necessary to compute another scale factor γ_i to transfer the local coordinate systems into a global unified one. As each image pair *i* and *j*, is part of at least two tuples (the ones with images *i* and *j* as reference, respectively) equation (4.14) can be formulated,

$$\frac{\gamma_i}{\gamma_j} = \frac{\eta_{ji}^{j}}{\eta_{ij}^{i}} = sf_{ij} \tag{4.14}$$

Employing the same rationale as in setting up equations (4.11) and (4.12), obtain

$$log\gamma_i - log\gamma_j = logsf_{ij} \tag{4.15}$$

$$\boldsymbol{A}_R \boldsymbol{X}_R = \boldsymbol{b}_R \tag{4.16}$$

where, $\boldsymbol{A}_R$, $\boldsymbol{X}_R$ and $\boldsymbol{b}_R$ are defined analogously to the terms in equation (4.12). Due to the rank defect of $\boldsymbol{A}_R$, γ_1 is set to 1. The scale factors γ of all tuples are simultaneously solved, again using the standard least square estimation:

$$arg \min_{X_R} \|\boldsymbol{A}_R \boldsymbol{X}_R - \boldsymbol{b}_R\|_2 \tag{4.17}$$

Finally, the global scale factor λ_{ij} of each pair can be determined by

$$\lambda_{ij} = \left(\gamma_i \eta_{ij}^{i} + \gamma_j \eta_{ji}^{j}\right)/2 \tag{4.18}$$

4.3.3 Solving global translations based on relative translations

Revisiting equation (4.8) after the scales are determined, only the image projection centers $\boldsymbol{C_i}$ remain unknown. Multiplying both sides of equation (4.8) by $\boldsymbol{R_i^{-1}}$, a linear equation system is eventually set up to determine the global translations,

$$(\boldsymbol{C_j} - \boldsymbol{C_i}) = \lambda_{ij} \boldsymbol{R_i^{-1}} \boldsymbol{t_{ij}} \tag{4.19}$$

To estimate all the projection centers, a linear equation system is generated by integrating equation (4.19) with all the edges in the viewgraph:

$$\boldsymbol{A}_P \boldsymbol{X}_P = \boldsymbol{b}_P \tag{4.20}$$

where $\boldsymbol{X}_P$ and $\boldsymbol{b}_P$ are vectors consisting of image projection center coordinates $\boldsymbol{C}_i$ and $\lambda_{ij} \boldsymbol{R_i^{-1}} \boldsymbol{t_{ij}}$, respectively. $\boldsymbol{A}_P$ is a sparse matrix in which three consecutive rows are all zeros, unless two corresponding images form an image pair which exists in the viewgraph. This matrix also has a rank deficiency due to a missing datum. The first image is defined as the origin of the global coordinate system to solve for that deficiency, i.e. $\boldsymbol{C_1} = \boldsymbol{0}$. Then, the coordinates of the projection centers of all images are estimated via standard least squares estimation (4.21),

$$arg \min_{X_P} \|\boldsymbol{A}_P \boldsymbol{X}_P - \boldsymbol{b}_P\|_2 \tag{4.21}$$

4.4 Robust bundle adjustment

After calculating the initial exterior orientation parameters of all available images, all the unknown object coordinates of the tie points are obtained by taking the average of multiple pairwise space intersections [Förstner and Wrobel, 2016, p.596]. As these initial values may contain errors and outliers exist due to spurious correspondences, the reconstructed tie points are checked and filtered using the condition that estimated tie point must always lie in front of all images in which they are visible. The final step of the global image orientation workflow, as Figure 1.2 illustrates, is the refinement by bundle adjustment. All the relevant initial results including orientation parameters and tie points' coordinates are optimized based on the functional model of the collinearity equations:

$$\begin{aligned} x &= -f\frac{r_{11}(X-X_0)+r_{21}(Y-Y_0)+r_{31}(Z-Z_0)}{r_{13}(X-X_0)+r_{23}(Y-Y_0)+r_{33}(Z-Z_0)}+x_0 \\ y &= -f\frac{r_{12}(X-X_0)+r_{22}(Y-Y_0)+r_{32}(Z-Z_0)}{r_{13}(X-X_0)+r_{23}(Y-Y_0)+r_{33}(Z-Z_0)}+y_0 \end{aligned} \tag{4.22}$$

with $\boldsymbol{R_i}=\begin{pmatrix} r_{11} & r_{21} & r_{31} \\ r_{12} & r_{22} & r_{32} \\ r_{13} & r_{23} & r_{33} \end{pmatrix}$

where (X, Y, Z) are the 3D coordinates of the *j-th* tie point X_j which is assumed to be viewed by the *i-th* image, and (x, y) are the corresponding 2D image coordinates $\boldsymbol{x_{ij}}$. (x_0, y_0) are the principal point coordinates of *i-th* image, f is its principal distance, (X_0, Y_0, Z_0) are the coordinates of the unknown projection center $\boldsymbol{C_i}$ (equivalent to the image translation vector), and r_{mn} (m =1, 2, 3; n =1, 2, 3) are the entries of the rotation matrix $\boldsymbol{R_i}$. Note that for the sake of simplicity, the indices i and j are omitted in equation (4.22).

The stochastic model is set up as an identity matrix for the covariance matrix, which assumes that all observations are uncorrelated and of equal uncertainty. Bundle adjustment is considered to be the maximum likelihood solution of the image orientation problem, but it relies on the quality the initial values. Since bundle adjustment is a standard optimization approach in photogrammetric procedures, and this thesis does not intend to contribute to this topic, only some necessary information on the implementation are introduced. More detailed information can be found in [Förstner and Wrobel, 2016].

Equation (4.22) is a typical representation of the collinearity function. However, the axis-angle representation is used in the implementation in this thesis, due to the fact that a representation with Euler angle may have problems with gimbal lock. Equation (4.23) is adopted to carry out the refinement task:

$$\underset{\boldsymbol{K_i},\boldsymbol{R_i},\boldsymbol{C_i},\boldsymbol{X_j}}{minimize}\sum_{i=1}^{N}\sum_{j=1}^{M}a_{ij}\left\|\boldsymbol{x_{ij}}-\varphi(\boldsymbol{K_i},\boldsymbol{R_i},\boldsymbol{C_i},\boldsymbol{X_j})\right\|_{huber} \tag{4.23}$$

where N is the number of images considered and M is the number of computed object points, while $a_{ij} = 1$ if an object point j is visible in an image i, otherwise, $a_{ij} = 0$. $\boldsymbol{K_i}$ contains the intrinsic parameters (x_0, y_0, f). $\boldsymbol{R_i}$, $\boldsymbol{C_i}$, $\boldsymbol{K_i}$ and $\boldsymbol{X_j}$ are considered as the items which need to be updated with improved estimations. Hence, a bundle adjustment with simplified self-calibration is performed (note that each image has its own set of intrinsic parameters, however, images taken from the same camera with the same setting have identical intrinsic parameters). The vector $\varphi = (x, y)^T$ contains the back projected x and y coordinates according to the collinearity equations (4.22), and $\boldsymbol{x}_{ij}$ are the observed 2D image coordinates. The Huber loss function (see Equation (4.24)) is used in the refinement pipeline, as it is less sensitive to observations with large residuals than the standard least squares estimation: the squared error loss ($0.5e^2$) is used only if the absolute value of residuals is smaller than 2 pixels, otherwise, $2(|e|-1)$ is used as loss function instead.

$$f(e) = \begin{cases} 0.5e^2 & if\ |e| \leq 2 \\ 2(|e| - 1), & otherwise \end{cases} \tag{4.24}$$

Equation (4.23) is iteratively optimized using formula (4.24) and the unknowns are updated after every iteration. Two criteria are adopted for termination: (a) the difference between the total cost value of equation (4.23) from the last to the current iteration is smaller than a threshold (ϵ_{ba}) times the current total cost value of equation (4.23); (b) the number of iterations is larger than T_{rba}. Erroneous object points occur rather often in SfM and have a negative effect on the final results. To obtain a more robust solution, two strategies are proposed to detect these errors: observations with residuals larger than a threshold v_r are eliminated as well as object points for which the corresponding largest intersection angle of all rays generating this point is smaller than a threshold d_a. To make it solvable, for each image in the photogrammetric block, a threshold T_{op} is introduced defining the minimum number of object points that need to be visible in an image. If an image does not have a sufficient number of observations, it is excluded from the block.

4.5 Discussion

This chapter provides a detailed explanation on how the exterior orientation parameters of the images in the filtered viewgraph $\mathcal{G}$ can be estimated in a global manner. Based on the characteristics of $SO(3)$, a coarse to refined global rotation estimation method is presented. More specifically, two robust estimators using L_1 norm and a Huber-like loss function are integrated to provide an efficient and robust solution for rotation estimation, identical to the approach presented by Chatterjee and Govindu [2013]. Next, a global linear translation estimation method is presented using relative translations and tie points concatenated to form tuples. Lastly, the estimated initial values are refined by robust bundle adjustment. Nevertheless, there are several inherent limitations which are worth further attention to gain a deeper understandings of the potential of the approach presented:

1) In this chapter, several least squares applications are introduced to solve various overdetermined problems, the main concentration is on the description of function model, and the related stochastic

models are not discussed. Although various strategies have been suggested to execute the global image orientation method in a robust fashion, it would be interesting to also investigate the stochastic model of each least square application. This would not only make the method more theoretically elegant, but also make a numerical reliability analysis possible for the estimated solution.

2) To remove the gauge datum when solving global translations using relative translations, the first image is always set as the reference which means its rotation and translation are fixed by an identity matrix and zero vector, respectively. While this is the most straightforward idea to solve the datum problem, it is not always the best choice. According to Wilson et al. [2016], fixing an image with a higher number of connected images can reveal more local convexity than using one from the periphery of the block. Therefore, it might be meaningful to investigate various choices for fixing the gauge datum, potentially providing a solution closer to the global optimum.

5 Experimental setup

In the last two chapters, the global image orientation approach with focus on improving time efficiency and robustness is described. The goal of the conducted experiments is to assess the performance of the proposed methods on various datasets. In particular, the time efficiency, robustness and accuracy of the preprocessing steps as well as the global image orientation method are investigated. For this purpose, this chapter is devoted to establishing the experimental setup for the subsequent evaluation. Section 5.1 introduces the objectives of the experiments which also reflect the goals of this thesis. According to the objectives, various datasets used in the experiments are introduced in section 5.2. The following section 5.3 provides more details on the implementations, such as information on the hardware and on free parameter settings being used. Finally, to achieve a convincing evaluation, section 5.4 presents the strategies and criteria that are adopted for the experimental evaluations.

5.1 Objectives of the designed experiments

In general, the primary objective of the experiments is to investigate the efficacy of the presented approach as a whole. Nevertheless, referring to Figure 1.2, the main developed components of the whole workflow are: detection of mutual overlapping image pairs, detection of RO outliers and global translation estimation. Therefore, in the evaluation phase three corresponding experimental objectives are designed to be investigated:

Objective 1: Assessment of the detection capability of mutually overlapping image pairs. To cope with a large number of images which are not regularly organized, in section 3.1, a time efficient image matching strategy based on a random k-d forest is suggested. Following this strategy, mutual overlapping image pairs are detected first, before homogenous points and the relative orientations are derived from them. As the corresponding results of this step are the input for the subsequent image orientation process, the first objective of the designed experiments is to validate the efficacy of the proposed approach and to investigate the corresponding time efficiency. More specifically, it is to be investigated that how many correct overlapping image pairs can be found as well as the amount of incorrectly detected overlapping image pairs. Furthermore, the computational time is evaluated.

Objective 2: Assessment of the robustification of the input ROs. As the original input, generated as described in Section 3.1, usually still contain gross errors, in Section 3.2 the idea of eliminating

these errors from the input ROs is introduced to improve the robustness for the subsequent image orientation estimation. First, based on the relative rotations and translations, a general method integrating the concept of triplet compatibility is presented; Second, different approaches are proposed to deal with blunders due to repetitive structure, very short baselines and baselines parallel to the viewing direction, respectively. Therefore, the second objective of the experiments is to demonstrate that the proposed RO outlier detection method is valid and can improve the performance of image orientation. In this context, the evaluation on dealing with images with repetitive structure and inappropriate baselines is of particular interest.

Objective 3: Assessment of global image orientation results. The final expected outputs of the whole thesis are the image orientation parameters and the coordinates of generated tie points, while the main corresponding contribution to global image orientation is the solution of global translation estimation. The previously estimated image orientation parameters and coordinates of tie points are refined by employing a robust bundle adjustment. Thus, the final objective of the experiments is to evaluate the image orientation results. For this purpose, the determined translations before and after bundle adjustment are evaluated by investigating the corresponding precision and accuracy. In addition, the processing time for carrying out image orientation is also taken into account.

5.2 Test datasets

To investigate the listed objectives and their fulfilment, various datasets are employed to perform the evaluation. Table 5.1 contains some general information about the experimental image datasets. Based on the organization of the images, there are in general two classes of datasets: *ordered* image sets which were acquired sequentially using an identical camera, and *unordered* image sets which were acquired in arbitrary ways using different cameras. This is also the reason that the image size of ordered data is constant per dataset, while in the unordered datasets the size may vary. All the used datasets are cited from the related publications. In this thesis, the essential matrix is used for computing the relative orientation; the intrinsic parameters, namely the principal point coordinates and principal distance, are either obtained from EXIF meta-information or from calibration information provided by the corresponding authors (only when this is necessary).

Which of the datasets are used to examine the individual objectives is listed in Table 5.1 as well. In general, *Objective 3*, assessing the performance of global image orientation, is investigated for all the datasets. Because the ultimate goal pursued in this thesis is to develop an image orientation method that can be applied to various datasets, the developed global image orientation method is tested on all categories of images. The related evaluation of the image orientation results is provided in the next chapter.

Category		Name of dataset	U or O	Number of Images	Image Size	Original	Intrinsic Parameters	GT	Objectives		
									1	2	3
ETH3D		*Facade*	O	76	6201×4130	Schops et al. [2017]	EXIF	Yes	✓	✓	✓✓
Oblique dataset		*3DOMcity*	O	420	6016×4016	Özdemir et al. [2019]	EXIF	Yes	✓	✓	✓✓
Internet datasets		Various (see Table 6.1)	U	Various (see Table 6.1)	Various	Wilson and Snavely [2014]	PbA	Yes	✓✓	✓✓	✓✓
		Various (see Table 6.2)	O	Various (see Table 6.2)	1936×1296	Olsson and Enqvist [2011]	EXIF	Yes	✓✓	✓✓	✓✓
Problematic Datasets	Benchmarks with RS and VSB ROs	*B1*	O	182	3936×2624	Wang et al. [2019c]	EXIF	Yes	✖	✓✓	✓✓
		B2		215					✖	✓✓	✓✓
		B3		342					✖	✓✓	✓✓
	Public datasets with high degree of RS	*TOH*	O	341	4368×2912	Shen et al. [2016]	EXIF	No	✖	✓✓	✓✓
		Sta.	O	156	4800×3200		EXIF	No	✖	✓✓	✓✓
		Ind.	O	152	1200×800		EXIF	No	✖	✓✓	✓✓
		Str.	O	175	3968×2232		EXIF	No	✖	✓✓	✓✓
		Capitole	O	99	3392×2264	Cohen et al. [2012]	EXIF	No	✖	✓✓	✓✓
		CAB	O	312	1696×1132		EXIF	No	✖	✓✓	✓✓
	Datasets with BPVD	*CR1*	O	33	2640×1760	Wang and Heipke [2020]	EXIF	No	✖	✓✓	✓✓
		CR2	O	24	2640×1760		EXIF	No	✖	✓✓	✓✓
		UAV1	O	48	4000×3000		EXIF	No	✖	✓✓	✓✓
		UAV2	O	47	4000×3000		EXIF	No	✖	✓✓	✓✓
	Complex dataset	*Church*	U	1455	Various	Michelini and Mayer [2020]	EXIF	No	✓	✓	✓✓
	Challenging dataset	*Quad*	U	6514	Various	Crandall et al. [2011]	PbA	Yes	✓	✓	✓✓

GT = Ground Truth; U or O = Unordered or Ordered; O = Ordered; U = Unordered; PbA = Provided by Authors

✖: The method of the corresponding objective is not used on the specific dataset.

✓: The method of the corresponding objective is used on the specific dataset, but the related detailed assessment is not given.

✓✓: The method of the corresponding objective is used, and the related detailed assessment is provided on the specific dataset.

Table 5.1: Datasets used in the experiments.

To demonstrate the proposed method's efficacy on detecting mutual overlapping image pairs (equivalent to *Objective 1*), two types of unordered and ordered internet datasets (see Figure 5.1 and 5.2 for sample images) are selected. However, in principal, this proposed method can also be applied on all the other datasets, but, as table 5.1 shows, it is not employed on the datasets with RS, VSB and BPVD. There are three reasons for not using the proposed time efficient image matching method: First, all these datasets are ordered with relatively small numbers of images, therefore the primary goal of establishing valid stereo pairs can be reached much easier. Second, all these datasets are captured in some deliberate way to obtain images with RS, VSB or BPVD. Thus, these datasets are particularly used in this thesis for conducting the evaluation of *Objective*

2. Figure 5.3 shows two sample images for each of these datasets. Those two sample images look very similar, but, they are in fact from two different real scenarios. Figure 5.4 shows samples of BPVD images which were captured at positions with various depths along the viewing direction. Lastly, when comparing against other methods, the same input should be used for a fair comparison. Those other methods, however, typically use exhaustive matching during preprocessing. The evaluation of *Objective 2,* eliminating outliers by checking the compatibility of triplets, is discussed for several datasets from ordered and unordered internet datasets, see table 5.1 the ground truth for the rotation and translation parameters is available as standard reference.

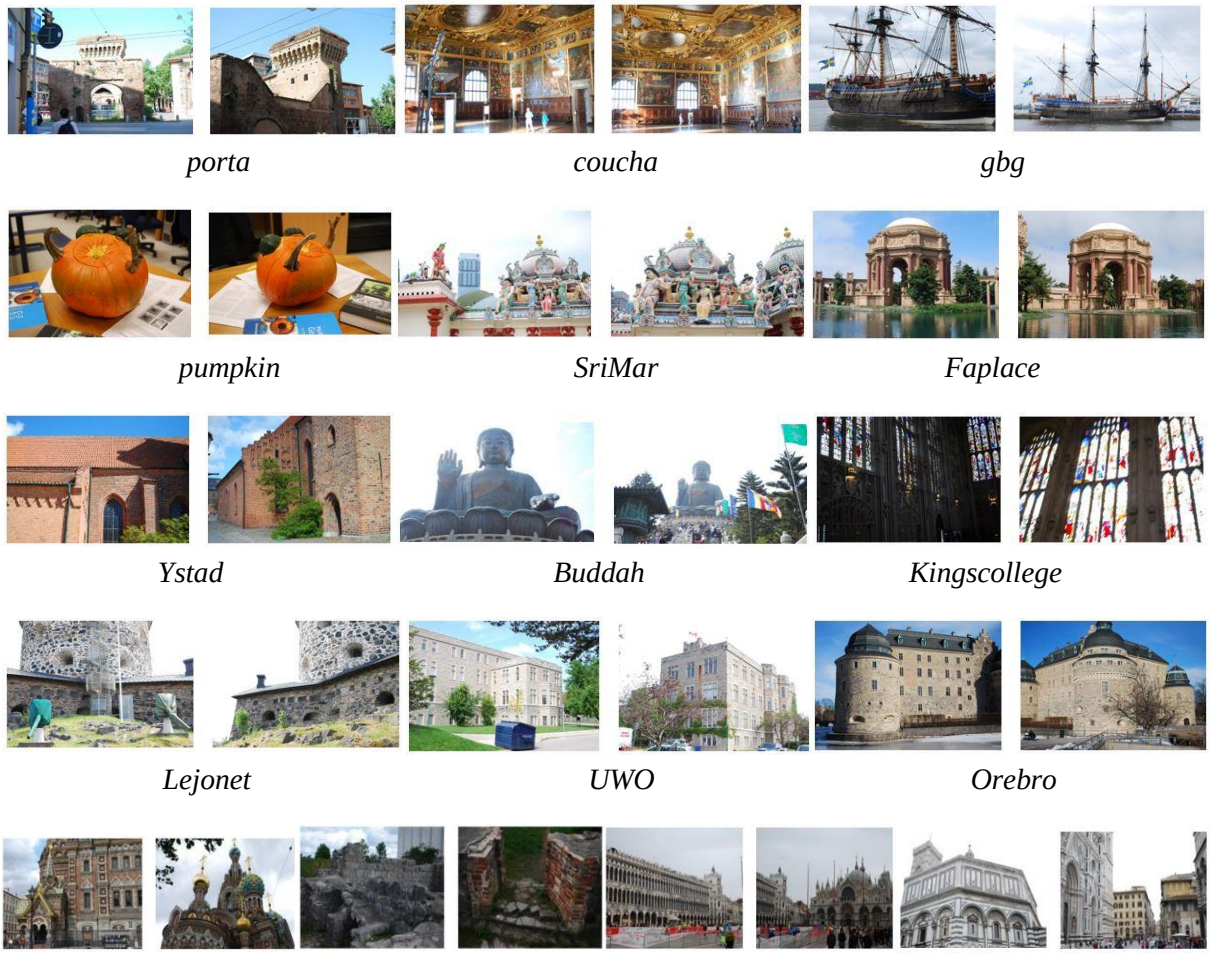


Figure 5.1: Sample images of ordered internet datasets.

To fully explore the potential of the proposed methods as a whole, the complex and the challenging dataset are investigated. The complex dataset consists of both UAV and terrain images leading to week pairwise epipolar geometric configurations. The challenging dataset contains different amount of image pairs with RS and VSB. Two sample images of both datasets are showed in Figure 5.5.

Ellis_Island *Piazza del Popolo* *NYC Library* *Metropolis*

York Minster *Montreal N.D.* *Tower of London* *Notre Dame*

Alamo *Gendarmenmarkt* *Vienna Cathedral* *Union Square*

Roman Forum *Piccadily*

Figure 5.2: Sample images of unordered internet datasets.

B1 *B2* *B3*

ToH *Sta.* *Ind.*

Str. *CAB* *Capitole*

Figure 5.3: Sample images of datasets with repetitive structure.

CR1 *CR2*

UAV1 *UAV2*

Figure 5.4: Sample images of BPVD datasets from positions with various depths.

Church *Quad*

Figure 5.5: Sample images of the complex and the challenging datasets, respectively.

5.3 Free parameter settings

There are a number of free parameters to be selected in the proposed methods (see Table 5.2). Their significance and selection are explained in the following. For image matching, those are *n_tr*, the number of k-d-trees in the forest, *d*, the minimum difference between nearest neighbouring features (expressed as the normalised scalar product between the two feature vectors) and cp_{min}, the minimum number of conjugate points per image pair. These parameters were investigated in Wang et al. [2017] and the values are chosen here according to the findings reported in that publication. Basically, *n_tr* is set to be 6 as a good compromise for time efficiency and the performance of detected mutual overlapping image pairs. While using more random trees can improve the result to a limited degree, the processing time increases significantly. Large values for cp_{min} will reduce the number of matched image pairs and some images may be excluded from the photogrammetric block. On the other hand, cp_{min} varies for different image matching methods, i.e., in the proposed random k-d forest based method only subsets of features are employed, while exhaustive pairwise image matching uses all features which aims to provide as many correspondences as possible for the RO calculation and tie point generation. The percentage of image pairs selected to be overlapping, *a*, is a rather sensitive parameter and is empirically selected based on some experiments conducted earlier and reported in Wang et al. [2019b]. In general, the photogrammetric block is extended if *a* is large, but at the same time, the block contains highly redundant ROs which may include blunders.

When verifying the epipolar geometric constraint, the number of RANSAC iterations (T_{rr}) is empirically selected. The maximum error in the epipolar geometry verification, v_e, is set to 4 pixels which is relatively large, its value takes into account the fact the intrinsic parameters of the sensors are only known with limited accuracy. To be considered as an inlier, an image pair must contain at least N_c conjugate points, which must account for more than b% of the total number per image pair.

To detect the RO outliers by checking the triplet rotation and translation compatibility, the two thresholds, ε_r and ε_t, are selected to be 5.0 degrees and 2.0, respectively, based on experience (note that ε_t should be proportional to the tuple consistent scale fixed by the reference image pair which is set to 1). These relatively large values again reflect the limited accuracy of the intrinsic sensor parameters; the more relaxed these two thresholds are set, the fewer ROs are typically detected as outliers, which results in a larger and denser photogrammetric block. The parameter *pt*, the minimum number of points per triplet for scale determination, is also selected heuristically.

		Name	Description	Value	Unit	Influence on PB
Image Matching	Efficient image matching based on random k-d forest	n_tr	Number of k-d trees	6	Tree	+
		d	Minimum distance to ensure neighbourhood between features (normalized scalar product of feature descriptor vectors)	0.7	∅	-
		cp_{min}	Minimum number of conjugate points per image pair	30	Point	-
		a	Percentage of image pairs selected as overlapping	35	∅	+
	Pairwise image matching	cp_{min}	Minimum number of conjugate points per image pair	80	Point	-
Epipolar Geometry Estimation		T_{rr}	Maximum number of RANSAC iterations in relative orientation	4096	iteration	?
		v_e	Maximum epipolar geometry error in relative orientation	4.0	Pixel	+
		N_c	Minimum number of conjugate points after epipolar geometry verification	50	Point	-
		b	Minimum percentage of conjugate points after epipolar geometry verification (from those before the check)	30	∅	-
ROs outlier detection		εr	Maximum allowed value of triplet rotation compatibility	5.0	Degree	+
		εt	Maximum allowed value of triplet translation compatibility	2.0	∅	+
Global translation estimation		pt	Minimum number of triplet points for scale transferring	5	Point	-
Robust Bundle Adjustment		ϵ_{ba}	Maximum change between unknowns from one iteration to the next to stop computations	10^{-6}	∅	+
		T_{rba}	Maximum number of iterations	50	iteration	-
		v_r	Maximum reprojection error	4.0	Pixel	+
		d_a	Minimum intersection angle	10.0	Degree	-
		T_{op}	Minimum number of object points observed by each image	15	Point	-

PB = Photogrammetric block
∅: parameter does not have a scale.
+: increasing the specific parameter value will result in a PB containing more images.
-: increasing the specific parameter value will result in a PB containing less images.
?: changing the specific parameter may have various influences on the PB.

Table 5.2: Parameter settings for the experiments.

For the bundle adjustment, two criteria ϵ_{ba} and T_{rba} are employed to terminate the optimization procedure. Specifically, the iterative refinement stops after T_{rba} iterations or if the total cost value (calculated by equation (4.23)) does not change from one iteration to the next one by more than ϵ_{ba} times the current total cost value. The refinement procedure would be terminated rather quickly if ϵ_{ba} is set to be very large value (e.g., 1) or if T_{rba} is selected to be a small value. In consequence, the orientation result of the photogrammetric block would be less accurate and less robust while more images might be retained in the block. To eliminate outliers, the maximum reprojection error v_r is set to be equal to the value v_e (which is 4 pixels) and, finally, the minimum intersection angle used to decide whether a point is accepted is set to 10 degrees. For each image, there should be at least some object points in the photogrammetric block, while the actual number

depends on the optimization procedure, e.g. more object points are needed if intrinsic parameters need to be refined. Thus, in this work, an image in the corresponding block must contain at least 15 object points (T_{op}).

Influence of parameter settings

It should be first noted that for all experiments identical parameter values (as suggested in the table) were used. Preliminary experiments suggested that their selection is not very critical. Nevertheless, in general, various selections do have an influence on the final PB up to some degree, as the last column 'Influence on PB' of Table 5.2 indicates. If a more complete and denser PB is desired (e.g., all captured images are required to be orientated for making sure that a complete reconstruction of the area can be obtained) parameters labelled with '+' may be increased and the ones marked with "-" may be reduced, and vice versa, if a sparse but more robust PB is desired.

5.4 Evaluation strategy and criteria

This section describes the strategy and criteria used in the experiments for the qualitative and quantitative evaluation of the proposed approaches. The corresponding strategy and criteria for assessing the preprocessing steps are discussed in section 5.4.1, which is followed by those for investigating the global image orientation in section 5.4.2.

5.4.1 Preprocessing steps

Detecting mutually overlapping image pairs

As Table 5.1 implies, the ordered and unordered internet datasets are used to assess the presented method for determining mutually overlapping image pairs. Since it is typically not easy to obtain the ground truth of real overlapping image pairs and one of the most widely used ways to solve this problem is to conduct exhaustive pairwise image matching, in this thesis the result of exhaustive pairwise matching is considered as the reference to evaluate the results of the detected overlapping image pairs. It worth to remind that this reference is not the best one and two noticeable weeknesses exist: some determined overlapping image pairs in the reference, in fact, may not overlap. This can resulte from some spurious matches, e.g., similar texture or limitations from the feature descriptor; On the other hand, some real overlapping image pairs may not be included by the reference, for example in the oblique aerial image set, some real overlapping image pairs between oblique images may not be found by using the conventional SIFT feature. Nevertheless, in the corresponding evaluation phase, the exhaustive pairwise image matching is still used as reference for the following reasons: first, the ground truth of real overlapping image pairs is not easy to generate, even a trained professional photogrammetrists can make a mistake when judging whether two images overlap or not and it is also infeasible to manually make such judgements for thousands of images. Second, the idea of exhaustive pairwise image matching is popular in many academic and commercial packages, e.g., *VisualSFM*, *Colmap*, *Photoscan*, and its efficacy has already been widely demonstrated.

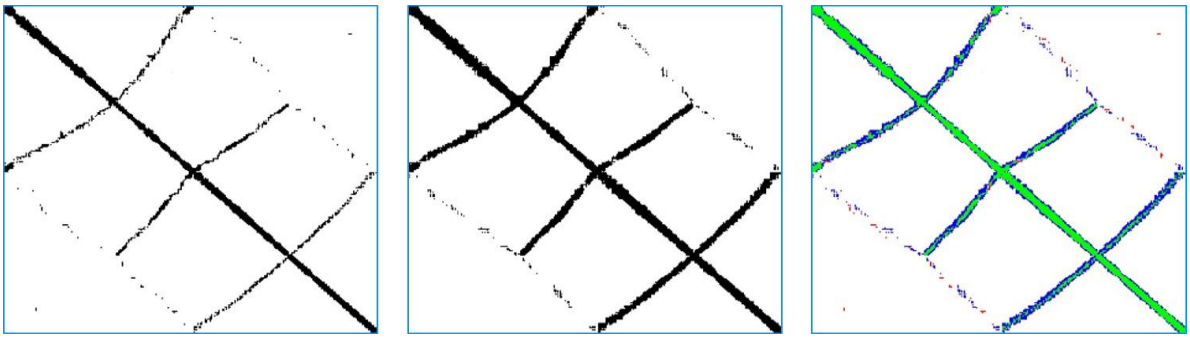
In addition, the results after epipolar geometry verification (some outliers of ROs and non-overlapping image pairs will be excluded) using exhaustive pairwise matchings are normally used as input for the subsequent image orientation by the mentioned packages. Therefore, based on this input, the epipolar geometric results of the detected overlapping image pairs are assessed as well. To allow for a quantitative evaluation, this problem can be treated as a classification issue which is similar to judging whether the two images of a detected overlapping pair really observe the same scene or not. Based on the estimated references, the criteria of *precision* and *recall* with and without epipolar geometry verification are studied. A generic confusion matrix is shown in Table 5.3, while the connection between the entries of the confusion matrix and the *precision* and *recall* are given in equation 5.1.

$$Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN} \tag{5.1}$$

	Actual positive	Actual negative
Predicted positive	*TP*	*FP*
Predicted negative	*FN*	*TN*

Table 5.3: Confusion matrix.

To give a more intuitive illustration, Figures 5.6 (a) and (b) show exemplary overlap graphs of the reference result by exhaustive pairwise matching and the detected overlapping pairs using the proposed method, respectively. The horizontal and vertical axes are the image IDs (Note that this also applies for all the relevant overlap graphs mentioned in the following contents), black indicates *overlap* and white *no overlap*. Figure 5.6 (c) is formed by taking the difference of the Figures 5.6 (a) and (b), pixels are labelled green (identical to *TP*) if the corresponding locations are both black, pixel are labelled blue (identical to *FN*) if Figure 5.6 (a) is white and Figure 5.6 (b) is black, and pixel are marked in red (identical to *FP*) if Figure 5.6 (a) is black and Figure 5.6 (b) is white, while white pixels existing in both Figures 5.6 (a) and (b) correspond to TN.



(a) Detected overlap graph (b) Reference results (c) Difference overlap graph

Figure 5.6: Example of overlap graph.

Lastly, the performance of one state-of-the-art method (called VocMatch [Havlena and Schindler, 2014]) is discussed. To compare the time efficiency, the running time of various methods are discussed as well.

Robustifying the ROs

In the evaluation of the method for making the input ROs more robust, the general approach for checking the compatibility of triplets and the individual approaches for coping with outliers due to RS & VSB and BPVD are inspected, respectively.

Method for checking triplets' compatibility. Two datasets, one ordered and one unordered, with ground truth exterior orientation parameters are tested, respectively. First, the ground truth relative rotations and relative translations are estimated from the ground truth exterior orientation parameters. Then, the discrepancies between the measured relative rotations and translations and the corresponding ground truth values are investigated (more details on computing these discrepancies can be found in Appendix B.1 and B.2). Among the detected RO inlier and outlier set, the graph of the cumulative distribution function (attributed with the relative rotation and translation errors) is employed to show the related performance. The desired performance is that the remaining ROs with small discrepancies should be mainly included in the inlier set, whereas, the ROs with large discrepancies should mainly be included in the outlier set.

	N_e	N_p	**Correct ROs**	**RS**	**VSB**
B1	182	2011	1089	784	138
B2	215	6357	1935	4030	392
B3	342	4956	3202	1422	332

Table 5.4: Three benchmarks with ROs ground truth.

Method for coping with ROs due to RS & VSB and BPVD. Analogous to Table 5.3, detecting RO outliers can also be cast as a classification problem if the ground truth of the ROs is known. In this context the term ground truth refers to the knowledge of each potential pair of images, whether it overlaps, whether it is non-overlapping with RS and whether it has a very short baseline. For the data sets *B1*, *B2* and *B3* of Table 5.1, the corresponding ground truth has been established manually by considering standard photogrammetric requirements. Some detailed information of these three benchmarks is listed in Table 5.4, where N_e is the largest number of connected images after filtering by the five-point algorithm and RANSAC, N_p indicates the corresponding number of ROs after this filtering step. The number of correct RS and VSB ROs for each dataset is also provided (note that there is no BPVD RO in these three datasets). Therefore, the criteria of the corresponding *precision* and *recall* values are again used to carry out the evaluation via a comparison with several RO outlier detection methods. In addition, the corresponding performance can also be indirectly revealed by comparing the image orientation results of using and not using the proposed methods (this strategy is in fact also applied on the other problematic datasets for which ground truth exterior orientation parameters are not available in the experimental parts), which is explained in the following image orientation evaluation stage.

Validation of the improved robustness on image orientation. To verify the hypothesis that the proposed methods for outlier detection are indeed capable of improving the image orientation results, two additional tests to investigate the quality of the estimated rotations and object points are carried out: *First, influence on global rotation estimation*. The basic idea of estimating global rotation matrices is to average the input relative rotations. Although this thesis employs the method

of Chatterjee and Govindu [2013] which is acknowledged to be relatively robust, it is interesting to see whether the robustified ROs can further benefit from this global rotation estimation and if so, to which degree the solution can be improved. In particular, the number of iterations and the processing time needed to converge are analyzed, and the accuracy of the global rotation solution is studied. *Second, influence on the reconstructed object point*. Based on the ETH3D facade dataset in which the ground truth of the object points is acquired using a laser scanning technique, the quality of the image orientation result is verified by inspecting various pipelines with and without the outlier detection in the stage of robustifying input ROs.

The completeness of the photogrammetric block

In the preprocessing steps both the procedures of detecting mutually overlapping image pairs and robustifying ROs can affect the size of the photogrammetric block, because these steps essentially determine whether the edges in the viewgraph can be maintained or need to be removed, while images with less than two connected edges will be eliminated from the block. Therefore, one crucial criterion for investigating the preprocessing steps is the corresponding value of $\boldsymbol{N_e}$ (largest number of connected images in the block) after each procedure.

5.4.2 Global image orientation

Basically, the evaluation of global image orientation is about the quality of the estimated exterior orientation parameters and the determined object points (which can only be evaluated if the related reference is available), and the corresponding runtime.

Figure 5.7: Top view of the simulated urban scenario.

First, a set of quasi-oblique aerial images [Özdemir et al., 2019] is tested. In a controlled lab environment over a 3D test field which simulates some common urban scenarios (as shown in Figure 5.7), 420 images (consisting of 144 nadir images and 276 oblique images) are captured, whereby the ground sampling distance (GSD) varies from 0.13mm to 0.27mm in the oblique images and amounts to 0.12mm in the nadir views. Three different criteria are suggested to evaluate the quality of the image orientation results: 1. Precision assessment: The reprojection errors of 115 targets labelled by red crosses in Figure 5.7 are reported to assess the precision of orientation results in image space. 2. Accuracy assessment: Three control bars (shown as blue lines in Figure 5.7) and three check bars (shown as yellow lines in Figure 5.7) with known length are

provided to evaluate the accuracy of the orientation results in object space. 3. Relative accuracy assessment: The errors of rotation and translation are inspected by taking the provided exterior pose parameters as a reference.

Then, the image orientation results of ordered and unordered internet datasets are assessed. As the global rotation solution is not the main contribution of this thesis, the assessment is mainly conducted on inspecting the accuracy of the global translation solution. More specifically, based on the ground truth exterior orientation parameters, the corresponding accuracies before and after the robust bundle adjustment are shown (note that the ground truth and the estimated image orientation results are given in different coordinate systems; see appendix C for transferring them into a unified frame and calculating the mean translation errors). To demonstrate that the presented global image orientation method is more time efficient compared to the conventional incremental method, one incremental approach by Wang et al. [2018] is tested on these datasets as well. In addition, several state-of-the-art global methods are compared in terms of the accuracy of translations and of the time efficiency, except one all investigated methods run on the same hardware.

Finally, the problematic datasets which typically contain critical stereo pairs are tested. As there is no ground truth for the exterior orientation parameters, qualitative results of the proposed global image orientation method are compared. The results of the proposed global image orientation method integrated with various RO outlier elimination methods are qualitatively investigated, and a coarse numerical analysis on the runtime and on the precision in image space is provided.

6 Evaluation

In this chapter, experimental results are presented and are thoroughly evaluated in order to clarify the advantages and limitations of the proposed methods. First, the performance of the proposed preprocessing steps is assessed (Sec. 6.1). In particular, the results of the methods to detect mutual overlapping image pairs and to increase the robustness of ROs are discussed. Afterwards, based on the inputs generated by these preprocessing steps, the evaluation of global image orientation (Sec. 6.2) is extensively investigated on various datasets (see Table 5.1). Finally, Section 6.3 closes this chapter with a synthesis of the reported experiments.

6.1 Evaluation of preprocessing steps

In this section, experiments conducted on parts of the listed datasets (see Table 5.1) are presented. The evaluation of the results for detecting mutual overlapping image pairs is carried out on all ordered and unordered internet datasets (Sec. 6.1.1), as they are assumed to be representative enough to demonstrate the capability of the presented method. Then, the efficacy of the proposed methods to increase the robustness of ROs are assessed (Sec. 6.1.2). Basically, the results are compared with the provided ground truth information (more details on the ground truth are given in Section 5.1).

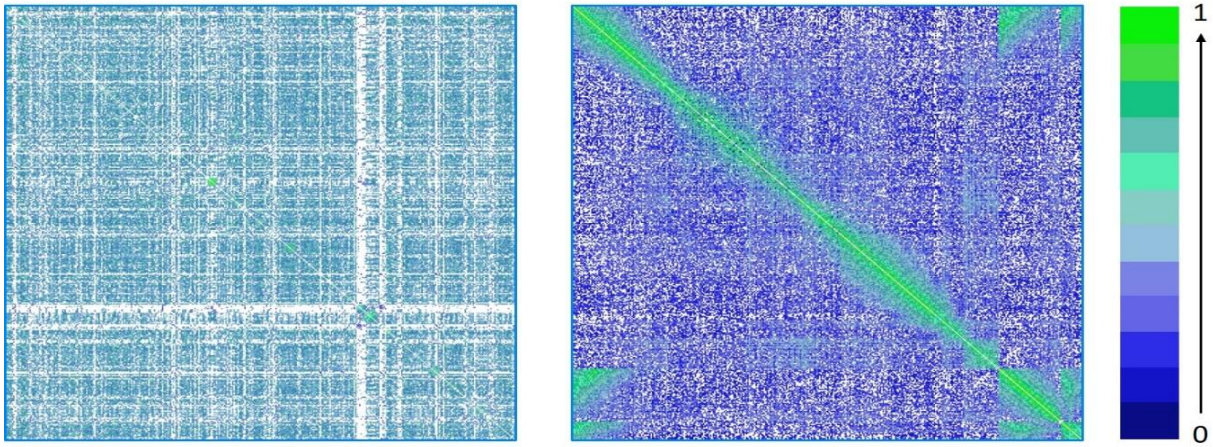
6.1.1 Performance of overlapping pair determination

Qualitative evaluation

To provide an intuitive impression of the results generated by the presented method to detect mutual overlapping image pairs, some qualitative results are first shown exemplarily for two datasets, one ordered internet dataset (*Lejonet*) and one unordered dataset (*Piazza del Popolo*), see Figure 5.1 and 5.2 for their sample images.

As the similarity degree value is crucial for the determination of final mutual overlapping image pairs, based on these two datasets, Figure 6.1 shows the corresponding overlap graphs attributed with their normalized similarity degree values (determined using a linear normalization). The colour of the pixel varies gradually from blue to green. While more bluish pixels indicate that the corresponding image pairs are supposed to have a lower probability to overlap, the more greenish pixels indicate that the corresponding image pairs are supposed to have a higher probability to overlap, the white pixels are the image pairs with the estimated S_{ij} equal to -1 which is due to that

less than 5 corresponding features are found. Comparing Figure 6.1(a) with Figure 6.1(b), one finding of the corresponding overlap graphs is that ordered datasets typically yield a more regular graph, while the overlap graph of an unordered dataset is usually irregular. The unordered images were captured in an arbitrary way which means that any image can potentially overlap with any other image. In contrast, the capturing process of ordered images is close to a sequential manner, which means that any image should overlap with its nearest neighbouring images (unless there is a closing loop). The efficacy of the proposed similarity degree value S_{ij} is revealed by Figure 6.1(b), because most pixels along the diagonal have been assigned with a greenish colour indicating that the corresponding images have a high probability to overlap with their neighbours (the green pixels on the right top and the left bottom corner are formed due to the fact that image capture occurred in a loop).

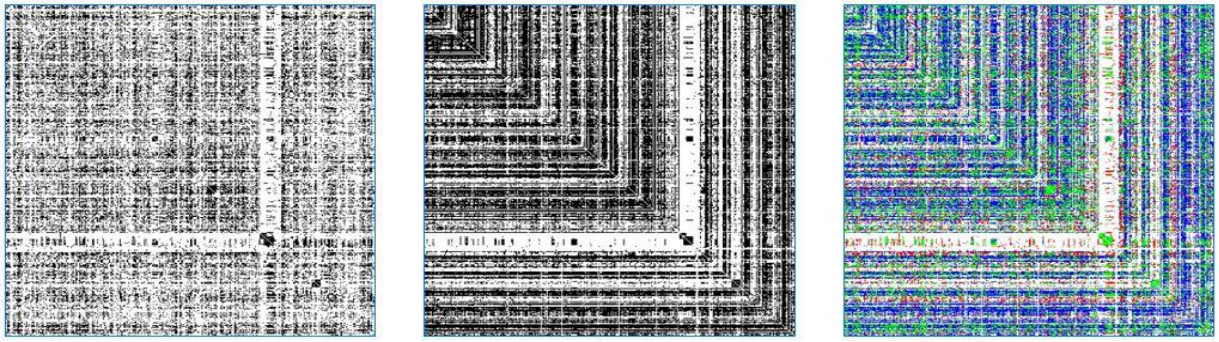


(a) *Piazza del Popolo* (b) *Lejonet*

Figure 6.1: Overlap graph attributed with normalized similarity degree value S_{ij}.

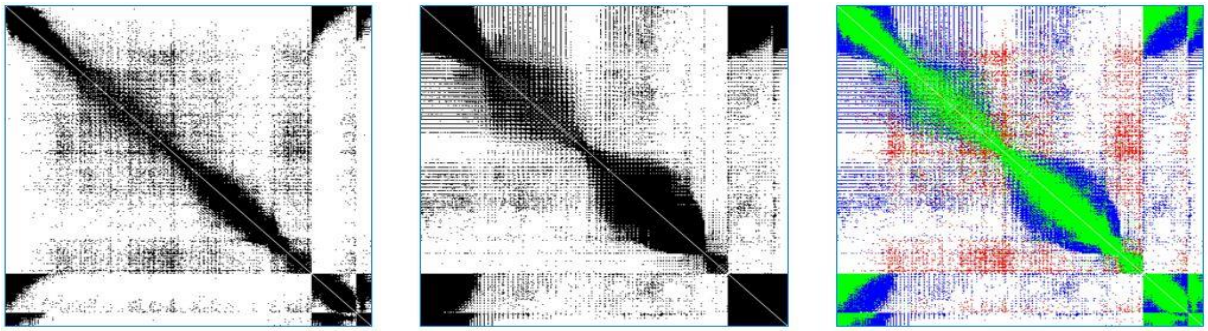
After computing the similarity degree value S_{ij}, image pairs that have the a% largest S_{ij} scores are chosen as potential overlapping pairs for a given image. Algorithm 3.1 is then adopted to refine these pairs by clustering images and discarding single images. Analogous to Figure 5.6, Figures 6.2(a) and 6.3(a) are the detected overlap graphs resulting from the proposed method, Figures 6.2(b) and 6.3(b) are the reference results using exhaustive pairwise image matching with the settings described in Table 5.2. Figures 6.2(c) and 6.3(c) are the difference graphs, when comparing the detected and reference overlap graphs. To qualitatively assess these results, the difference overlap graphs are produced (the meaning of blue, red, and green pixels is equal to the one in Figure 5.6, namely, pixel is labelled green which is identical to *true positive* if detected overlap graph and reference result are both black, pixel is labelled blue which is identical to *false negative* if detected overlap graph is white and reference result is black, and pixel is marked in red which is identical to *FP* if detected overlap graph is black and reference result is white) and imply that the detected overlapping image pairs are generally consistent with the reference as a whole. However, there exist some red pixels indicating that some predictions are actually incorrect and some blue pixels which means that some actually overlapping pairs are not successfully predicted. This can be explained by the fact that the reference is not the most rigorous ground truth and two weeknesses exist: some determined overlapping image pairs in the reference, in fact, do not

overlap; some real overlapping image pairs may not be included in the reference (see Sec. 5.4.1 for more details).



(a) Detected overlap graph (b) Reference result (c) Difference overlap graph

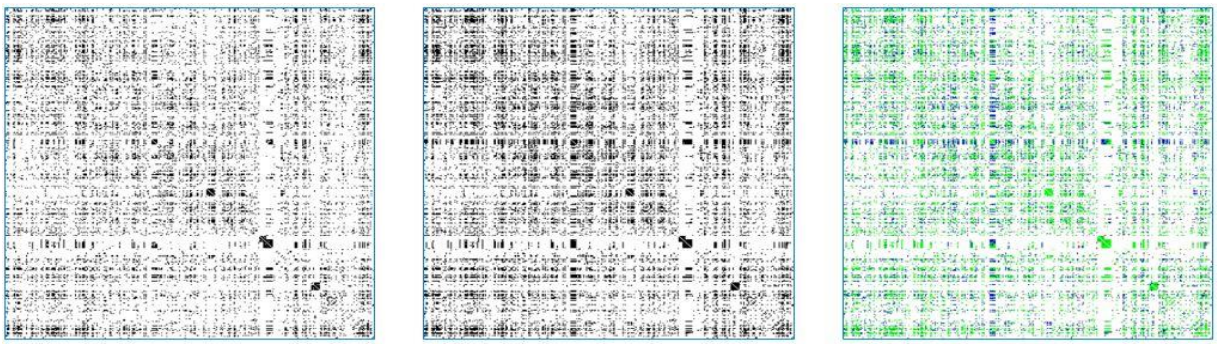
Figure 6.2: Overlap graph of Piazza del Popolo.



(a) Detected overlap graph (b) Reference result (c) Difference overlap graph

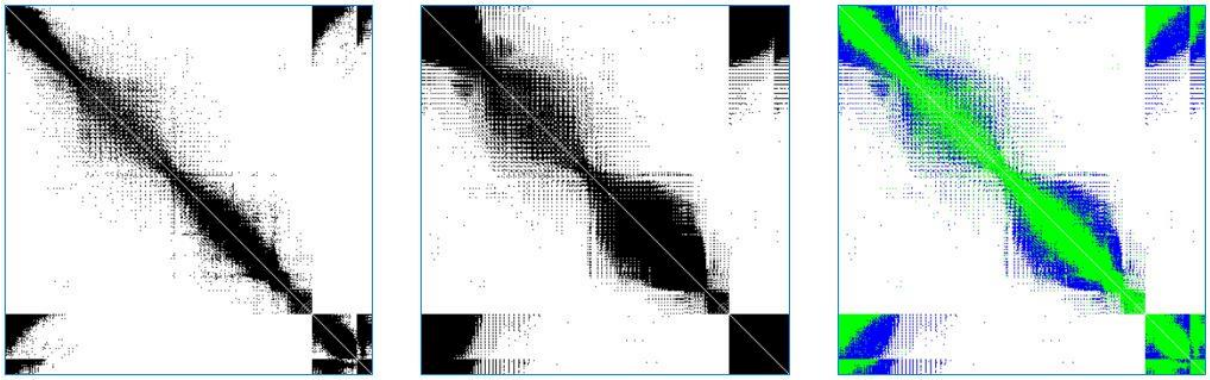
Figure 6.3: Overlap graph of Lejonet.

Figures 6.4 and 6.5 are the overlap graphs after filtering for epipolar geometric verification. It can be seen that in both datasets many overlapping image pairs detected by the proposed method and the exhaustive image matching method are eliminated as outliers, and that the difference overlap graphs are improved significantly: most of the original red pixels disappear and the number of blue pixels is reduced. This is due to the fact that image pairs incorrectly considered as overlapping in exhaustive matching and by the proposed method do not pass the epipolar geometric check and are thus discarded. On the other hand, the proposed method is able to determine a set of overlapping image pairs which have a high chance to pass the epipolar geometric check.



(a) Detected overlap graph (b) Reference result (c) Difference overlap graph

Figure 6.4: Overlap graph of Piazza del Popolo after epipolar geometric verification.



(a) Detected overlap graph (b) Reference result (c) Difference overlap graph

Figure 6.3: Overlap graph of Lejonet after epipolar geometric verification.

Quantitative evaluation

The qualitative evaluation illustrates the capability of the proposed method to some extent. Nevertheless, a quantitative evaluation is obviously needed for a more complete analysis of the proposed method. According to the evaluation strategy and criteria described in Section 5.4.1, the precision and recall of each ordered and unordered internet dataset are investigated and the corresponding runtime is reported. In addition, the results are compared against one state-of-the-art method, VocMatch [Havlena and Schindler, 2014].

Tables 6.1 and 6.2 show the numerical results for all unordered and ordered internet datasets investigated in this section, sorted by the number of images per dataset. Before running epipolar geometric verification, for the proposed method all precision values are higher than 0.7, which indicates that more than 70% of the found overlapping image pairs are identical to the reference determined by exhaustive pairwise matching. In addition, the precision values of ordered internet datasets are generally higher than those of unordered internet datasets, e.g., the lowest precision on ordered internet datasets is 0.85, and only 0.73 for unordered ones. The reason can be explored by revisiting Figure 6.1: the similarity degree value S_{ij} is more distinguishable for ordered internet datasets than for unordered ones. This clear difference is due to the fact that ordered image sets were taken in a sequential way, so neighboring images typically have similar content while images taken at an earlier or later epoch tend to depict different scenes. Thus, the S_{ij} values of these pairs are expected to be very small. However, unordered images typically show the same object from arbitrary viewpoints and many of these image pairs actually mutually overlap to various degrees, thus, the S_{ij} values of these pairs can vary a lot. In particular, the distribution of normalized S_{ij} values on ordered datasets is more close to be bi-modal, whereas, for unordered datasets the normalized S_{ij} values are more equally distributed with values ranging around the median values (0.4 – 0.7). The better the S_{ij} values can be separated in practice, the more beneficial they are for the task of classifying non-overlapping and overlapping image pairs. For each image, only the image pairs with the 35% largest similarity degree values are further considered. This explains the somewhat small recall values compared to the precision. Nevertheless, except for very few datasets (*Piccadily, Pumpkin, UWO, SanMarc*), the recall values are only higher than 0.4. A comparison

with VocMatch[1] [Havlena and Schindler, 2014] is shown in these two tables as well. To allow for a fair comparison, some parameters of VocMatch were adjusted to be identical to the ones used in this thesis (see Table 5.2). Considering the resulting adjacency matrix of VocMatch only, the proposed method outperforms VocMatch with respect to both precision and recall. Nearly all measures for both groups of datasets are larger for the proposed method compared to VocMatch, before and after epipolar geometric verificaiton.

Name	N	Before EG verification							After EG verification						
		Exh	PM			VocMatch			Exh	PM			VocMatch		
		N_e	N_e	P	R	N_e	P	R	N_e	N_e	P	R	N_e	P	R
Ellis_Island	247	247	245	0.88	0.52	233	0.83	0.37	243	239	0.96	0.59	231	0.90	0.52
Piazza del Popolo	354	354	345	0.80	0.45	343	0.74	0.34	348	335	0.94	0.52	343	0.89	0.49
NYC Library	379	379	371	0.73	0.49	356	0.64	0.35	372	363	0.93	0.69	351	0.91	0.52
Metropolis	394	393	388	0.74	0.42	385	0.60	0.21	388	378	0.95	0.53	385	0.88	0.50
York Minster	458	458	455	0.72	0.49	455	0.62	0.38	453	448	0.95	0.56	455	0.90	0.44
Montreal N.D.	474	474	468	0.76	0.51	468	0.66	0.40	471	464	0.95	0.60	461	0.92	0.51
Tower of London	508	508	506	0.73	0.54	486	0.68	0.39	499	489	0.93	0.62	479	0.87	0.54
Notre Dame	553	553	553	0.94	0.45	551	0.86	0.36	553	553	0.97	0.51	543	0.95	0.48
Alamo	627	620	601	0.92	0.41	611	0.75	0.35	611	579	0.99	0.58	605	0.86	0.50
Gendarmenmarkt	742	740	731	0.74	0.44	441	0.66	0.27	712	708	0.91	0.50	441	0.93	0.41
Vienna Cathedral	918	916	911	0.81	0.42	896	0.65	0.32	901	896	0.96	0.54	895	0.89	0.44
Union Square	930	930	909	0.76	0.48	901	0.70	0.36	899	879	0.97	0.55	901	0.94	0.46
Roman Forum	1134	1133	1131	0.77	0.46	1079	0.59	0.26	1112	1112	0.90	0.54	1103	0.86	0.43
Piccadily	2508	2416	2393	0.76	0.35	2221	0.71	0.29	2356	2311	0.86	0.44	2194	0.88	0.38

EG = Epipolar geometry; Exh = Exhaustive pairwise image matching; PM = Proposed method; P = Precision; R= Recall;
N = the number of images of each input dataset; N_e = the largest number of connected image in the photogrammetric block;
These abbreviations are also used in the subsequent tables in this section.

Table 6.1: Evaluation of image matching accuracy on unordered internet dataset.

Name	N	Before EG verification							After EG verification						
		Exh	PM			VocMatch			Exh	PM			VocMatch		
		N_e	N_e	P	R	N_e	P	R	N_e	N_e	P	R	N_e	P	R
Porta	141	141	141	0.91	0.41	141	0.92	0.33	141	141	1.0	0.54	141	1.0	0.48
CouCha	176	176	176	0.86	0.43	176	0.84	0.35	176	176	1.0	0.60	176	0.98	0.51
Gbg	179	179	179	0.92	0.49	179	0.88	0.36	179	179	1.0	0.53	179	1.0	0.47
Pumpkin	209	209	209	0.89	0.35	209	0.87	0.33	209	209	0.99	0.58	209	0.95	0.48
SriMar	222	222	222	0.94	0.42	222	0.91	0.42	222	222	1.0	0.54	222	1.0	0.52
Fapalace	281	281	281	0.93	0.41	281	0.92	0.34	281	281	1.0	0.50	281	1.0	0.48
Ystad	290	290	290	0.85	0.45	290	0.89	0.44	290	290	1.0	0.53	290	1.0	0.56
Buddah	322	322	322	0.92	0.44	322	0.94	0.39	322	322	0.97	0.51	322	1.0	0.48
Kingscollege	361	361	361	0.91	0.52	361	0.87	0.47	361	361	1.0	0.63	361	1.0	0.60
Lejonet	368	368	368	0.88	0.54	368	0.86	0.48	368	368	1.0	0.61	368	1.0	0.59
UWO	692	692	692	0.93	0.39	692	0.87	0.34	692	692	1.0	0.53	692	1.0	0.51
Orebro	763	763	763	0.94	0.41	763	0.94	0.30	763	763	1.0	0.54	763	10	0.50
Spilled	781	781	781	0.86	0.40	781	0.86	0.34	781	781	0.94	0.58	781	0.92	0.52
Ahus	811	811	811	0.87	0.42	811	0.85	0.38	811	811	0.99	0.55	811	0.99	0.54
SanMarc	1499	1499	1499	0.92	0.36	1499	0.90	0.28	1499	1499	1.0	0.49	1499	1.0	0.36
Duomo	1805	1805	1805	0.85	0.41	1805	0.84	0.27	1805	1805	0.98	0.52	1805	0.96	0.43

Table 6.2: Evaluation of image matching accuracy on ordered internet dataset.

After epipolar geometric verification, for both methods, the proposed method and VocMatch, precision and recall on both ordered and unordered datasets are improved, which is consistent with the observations on the qualitative results. On the one hand, more than 90 percent of detected

[1] VocMatch can be found at https://www.ethz.ch/content/dam/ethz/special-interest/baug/igp/photogrammetry-remote-sensing-dam/documents/sourcecode-and- datasets/Vocmatch/vocmatch-1.0.zip

overlapping image pairs by the proposed method are correct. On the other hand, over 50 percent of the reference image pairs are successfully detected. Special attention should be paid to low recall values, because this may indicate that the photogrammetric block is not reconstructed completely. For this purpose, the information of N_e after each procedure is shown in Tables 6.1 and 6.2. For the unordered internet datasets, few images which are weekly connected to the photogrammetric block are excluded after exhaustive matching and EG verification. Compared with the reference, the proposed method keeps almost the same number (for some datasets just slightly less) of images. For the ordered internet datasets, the proposed method and VocMatch provide exactly the same N_e value as the exhaustive pairwise matching.

	Name	Exh			Proposed Method				VocMatch
		Matching	EG	Total	RKF	Matching	EG	Total	Total
U	*Ellis_Island*	899	135	1034	28	330	47	**405** (×2.6)	457
	Piazza del Popolo	1429	200	1629	37	324	51	**412** (×4.0)	483
	NYC Library	2126	236	2362	32	616	79	727 (×3.2)	**569**
	Metropolis	1790	218	2008	30	381	26	**437** (×4.6)	477
	York Minster	3932	440	4372	39	489	64	**592** (×7.4)	619
	Montreal N.D.	4865	506	5371	46	870	113	1029(×5.2)	**880**
	Tower of London	3555	411	3966	48	322	43	**413** (×9.6)	503
	Notre Dame	10663	1034	11697	50	1219	166	**1435**(×8.2)	1443
	Alamo	4998	662	5660	71	562	80	713 (×7.9)	**661**
	Gendarmenmarkt	5558	1015	6573	78	714	118	**910** (×7.2)	1046
	Vienna Cathedral	17664	2044	19708	101	1167	154	1422(×13.8)	**1265**
	Union Square	13300	1913	15213	101	979	107	**1187**(×12.8)	1388
	Roman Forum	23582	3685	27267	128	923	88	1139(×15.0)	**1130**
	Piccadily	130216	16735	146951	380	2706	468	3554(×41.3)	**3332**
O	*Porta*	497	75	572	33	124	40	**197** (×2.9)	211
	CouCha	1169	157	1326	42	332	79	**453** (×2.9)	489
	Gbg	636	123	759	38	175	45	**258** (×2.9)	303
	Pumpkin	321	66	387	39	125	26	190 (×2.0)	**177**
	SriMar	1189	203	1392	51	313	73	**473**(×2.9)	512
	Fapalace	1686	323	2009	66	416	148	**630** (×3.2)	689
	Ystad	4002	554	4556	84	529	156	769 (×5.9)	**734**
	Buddah	801	224	1025	60	251	71	**382** (×2.7)	369
	Kingscollege	2580	432	3012	87	788	161	1036 (×2.9)	**986**
	Lejonet	5915	634	6549	97	816	214	**1127** (×5.8)	1444
	UWO	15957	1319	17276	213	1237	255	**1705**(×10.1)	2010
	Orebro	16295	2136	19061	253	1835	427	2515 (×7.6)	**2469**
	Spilled	16973	2897	19870	244	1777	404	**2425** (×8.2)	2845
	Ahus	21810	1487	23297	258	1329	219	**1806**(×13.2)	2054
	SanMarc	59149	11206	70355	581	2376	848	3806(×18.2)	**3741**
	Duomo	118036	20040	138076	767	3536	1123	**5446**(×25.4)	5712

U = Unordered internet dataset; O = Ordered internet dataset; RKF = random k-d forest;

Table 6.3: Runtime in seconds for all experiments on the internet datasets. The factor in brackets indicates the speed-up of the proposed method compared to exhaustive matching. The bold font indicates the fastest method.

Table 6.3 lists the runtime of the various steps for image matching and performing epipolar geometric verification. All of these experiments are executed on the same machine with a quad-core processor (3.2 GHz Inter (R) Core (TM)i5-6500, 32G memory) and eight threads in total. Compared with the conventional exhaustive pairwise matching, the suggested image matching

strategy only considers the determined overlapping image pairs. This improves the speed by a factor between 2.6 to 41.3 (see Table 6.3) depending on the size of the dataset, while the factor generally increases as the number of input images grows (see *N* of each dataset in Table 6.1 and 6.2). It is clear that exhaustive image matching needs more time for more images, because more features and more potential pairs are needed to be dealt with. This is also true for the runtime of executing the random k-d forest, as more time is required to build a random k-d forest for more images. When comparing the proposed method and VocMatch, similar runtimes can be observed for both method (both show a runtime in the same order of magnitude, in contrast to exhaustive pairwise image matching) to complete the image matching and EG verification task. Note that absolute runtimes from different datasets are hard to compare, as the number of images are different per dataset, furthermore, for the unordered image sets the image size varies considerably even within the same dataset (this can be found by inspecting the original image sets[2]) which can result in that the number of extracted features might vary greatly.

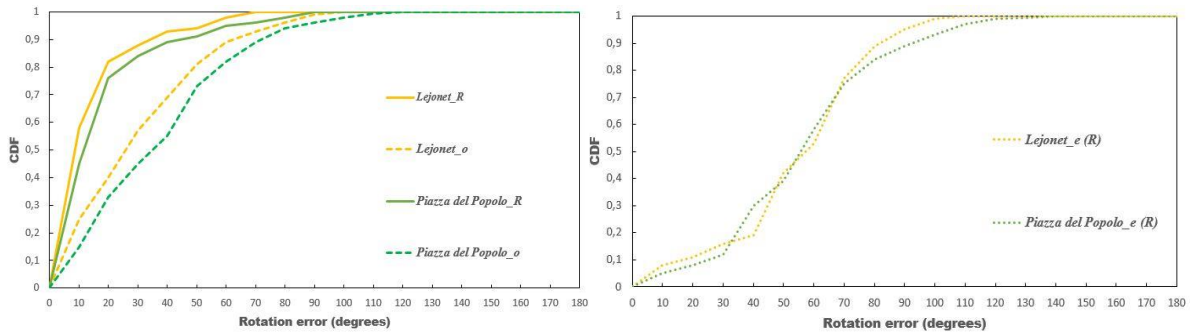
6.1.2 Performance of the robustification of ROs

In this section, the evaluation of the proposed general approach to increase the robustness of ROs by checking the triplet compatibility is first studied on one ordered and one unordered internet dataset (*Lejonet* and *Piazza del Popolo*). Then, the detection of outlier ROs due to RS and VSB is analysed on the benchmarks specified by the relevant objective (see Table 5.1) for which the ground truth ROs are provided. Finally, the image orientation results which are obtained using and not using the robustified input ROs are compared.

Robust ROs by checking triplet compatibility

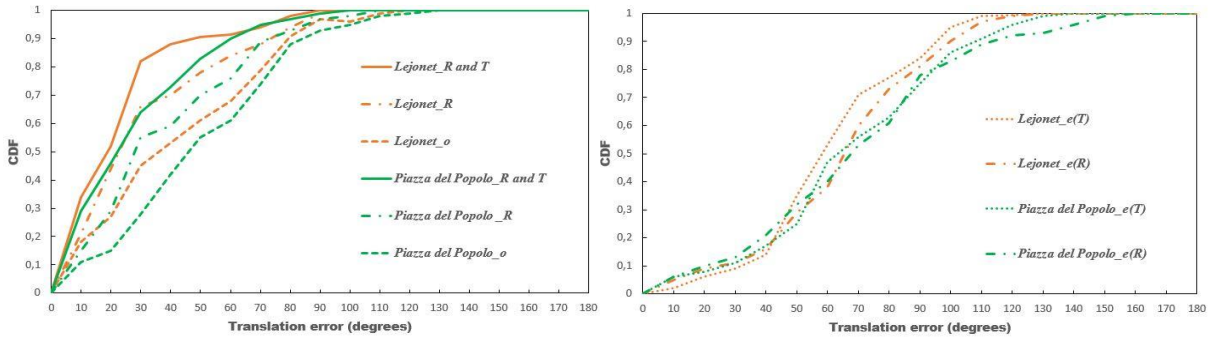
First, the relative rotations before and after rotation outlier detection are evaluated by comparing them with the ground truth (see Section 5.4.1). Figure 6.6 shows the results in terms of the cumulative distribution function of the different variants compared to ground truth. The arc cosine of the average value of the main diagonal elements from the difference rotation matrix between the ground truth and observed relative rotation matrix is used (see Appendix B.1). As Figure 6.6(a) shows, before carrying out rotation outlier detection, only 33% of the differences are smaller than 20 degrees for *Piazza del Popolo*, and only 40% for *Lejonet*. After outlier detection, nearly 80% of the rotations show a difference from the ground truth below 20 degrees (76% and 82% for *Piazza del Popolo* and *Lejonet*, respectively). While this demonstrates that the rotations are inaccurate before performing rotation averaging, it is a clear indication that the proposed rotation outlier detection approach is very useful. To see how many good relative rotations are eliminated, Figure 6.6(b) is shown. Among these eliminated relative rotations, only less than 10% have errors below 20 degrees, and most of the eliminated relative rotations' errors are between 40 and 100 degrees.

[2] The original unordered image sets can be accessed by http://www.cs.cornell.edu/projects/1dsfm/



(a) (b)

Figure 6.6: The cumulative distribution function (CDF) of relative rotation errors. In (a), "Lejonet_R" is the result after outlier detection by checking the rotation triplet compatibility, "Lejonet_o" denotes the result without using the corresponding outlier detection method. For the other dataset, the same notation is applied. In (b), "Lejonet _e(R)" is the result of eliminated relative rotations by checking the rotation triplet compatibility, the same notation is applied for the other dataset.



(a) (b)

Figure 6.7: The CDF of relative translation error. In (a), "Lejonet_R and T" is the result after outlier detection by combing the check of both rotation and translation triplet compatibility, "Lejonet_R" is the result after outlier detection by only using rotation triplet compatibility, "Lejonet _o" denotes to the result without using any outlier detection. For the other dataset, the same notation is applied. In (b), "Lejonet _e(R)" is the result of eliminated relative translations by using rotation triplet compatibility and "Lejonet _e(T)" is the result of eliminated relative rotations by using translation triplet compatibility. The same notation is applied on the other dataset.

A similar comparison was carried out for the combined outlier detection in relative rotation and translation. The intersection angle between the derived relative translation vector and the ground truth is used as the evaluation criterion (see Appendix B.2). Figure 6.7(a) shows, that before outlier detection, only 28% of the translation directions are within 30 degrees of the ground truth for *Piazza del Popolo*, and only 45% for *Lejonet*. After rotation outlier detection according to rotational triplet compatibility, the corresponding results for *Piazza del Popolo* increase to 55% and for *Lejonet* to 66%. This indicates that wrong rotation and translation values are indeed highly correlated (see Section 3.2.1). Performing also the additional outlier detection according to translation triplet compatibility yields values in the range of 64% for *Piazza del Popolo*, and 82% for *Lejonet.* Again, while these values are still somewhat inaccurate, the benefit of our outlier detection approach is clearly visible. Figure 6.7(b) shows the CDF of the eliminated relative translations using rotation and translation triplet compatibility, respectively. The trends of the shown curves are very similar to those of Figure 6.6 (b): for both rotation and translation compatibility checking, less than 10% of the eliminated relative translations have an error below

20 degrees, while most of the eliminated relative translations have errors between 30 and 100 degrees.

Robustification of ROs due to RS and VSB

To validate the performance of the corresponding presented method to detect incorrect relative orientations (ROs) due to repetitive structure (RS) and very short baseline (VSB), Table 6.4 provides the obtained precision and recall values for detecting RS, VSB and the correct ROs. For this purpose, the benchmarks with RS and VSB ROs (see Table 5.1) and the relevant ground truth ROs (see Table 5.4) are used. It can be observed that most ground truth ROs can be detected (recall is higher than 90%), while also many of our precision values are higher than 90% (with B2 and B3 reaching 89.6% and 88.6% for VSB only, respectively). These results illustrate that most of the identified ROs are correctly classified as correct, RS or VSB by the presented method.

	Detection of RS only		**Detection of VSB only**		**Correct ROs after RS and VSB elimination**	
	P	R	P	R	P	R
B1	92.7	90.2	90.6	92.4	90.2	97.0
B2	95.2	91.2	89.6	90.6	95.5	96.0
B3	91.4	90.2	88.6	93.2	93.3	92.6

Table 6.4: Precision and recall values in percent on detecting RS, VSB and correct ROs. P and R denote precision and recall.

	Ground truth	**PM_RSVSB**	**PM_GTC**	**Wang et al. [2019c]**	**Wilson and Snavely [2014]**	**Zach et al. [2010]**
B1	1089	1171	1846	1303	1569	1684
B2	1935	1946	5066	1918	5391	5839
B3	3202	3178	4776	3278	3690	4349
PM_RSVSB = The presented method for detecting outlier ROs due to RS and VSB (see Section 3.2.2 and 3.2.3); PM_GTC = The presented method for detecting outlier ROs using the general triplet compatibility check (see Section 3.2.1); These two abbreviations are used in subsequent tables, figures and contexts as well.						

Table 6.5: Comparison of the number of selected ROs from different methods.

	PM_RSVSB		**PM_GTC**		**Wang et al. [2019c]**		**Wilson and Snavely [2014]**		**Zach et al. [2010]**	
	P	R	P	R	P	R	P	R	P	R
B1	**90.2**	97.0	56.4	95.2	81.4	**97.4**	65.2	94.2	59.8	92.2
B2	**95.5**	96.0	40.2	**98.7**	93.8	93.0	37.3	98.6	35.3	**98.7**
B3	**93.3**	92.6	66.5	**99.1**	91.3	93.5	81.9	95.2	73.1	98.8

Table 6.6: Comparison of precision and recall value in percent of different methods. P and R denote precision and recall; the best values are highlighted.

To further investigate the presented RS and VSB ROs elimination method, the results of PM_GTC (denoted as the proposed method only using the general triplet compatibility check), Wang et al. [2019c], Wilson and Snavely [2014] and Zach et al. [2010] are compared. Among these methods, Wang et al. [2019c] suggested the same criteria as this thesis introduced for detecting RS and VSB ROs, but for different datasets various settings of free parameters have to be selected in advance for good results (note that here the results of Wang et al. [2019c] with individually refined free parameters are presented), Wilson and Snavely [2014] detected RO outliers by inspecting the geometric inconsistencies of the relative translations, [Zach et al, 2010]

presented a method which adopted the loop consistency constraint to infer the invalid ROs. Table 6.5 shows that the number of selected ROs for Wang et al. [2019c] and the PM_RSVSB (denoted as the proposed method by detecting and eliminating RO outliers due to both RS and VSB) are the two smallest ones (and closest to the ground truth values from Table 5.4); this is also illustrated by Figure 6.8 in which the overlap graphs of PM_RSVSB and Wang et al. [2019c] are filled with less black pixels than those of the other methods, more RO outliers are eliminated by these two methods. Based on the ground truth ROs, the precision and recall values are estimated (by employing Equation 5.1), as Table 6.6 shows. It can be seen that the recall values of all methods are higher than 90%, which means they are all able to detect most of the correct ROs, whereas the PM_RSVSB clearly outperforms the others in terms of precision.

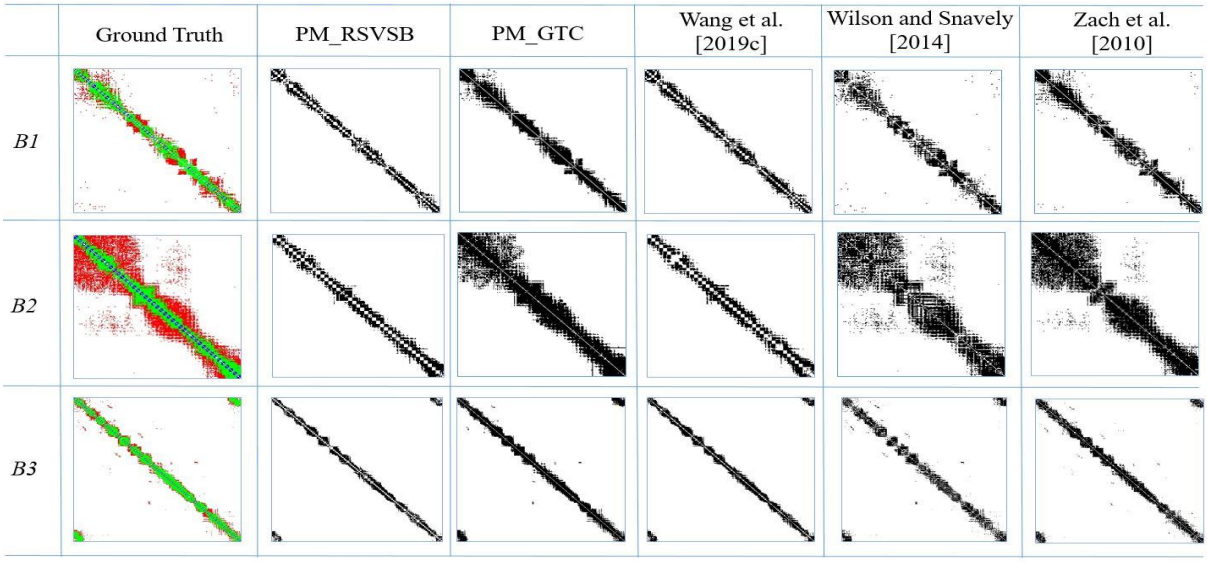


Figure 6.8: Overlap graphs of the three benchmark datasets from different methods. The second column is the RO ground truth, where green pixels denote correct ROs, red pixels are RS ROs and blue pixels denote VSB ROs. In the following five columns, black pixels indicate that the corresponding ROs are kept and white pixels represent non-overlapping image pairs.

Some additional experiments on six public datasets with highly repetitive structure, namely the Temple of Heaven (*ToH*), Indoor (*Ind.*), Stadium (*Sta.*), Street (*Str.*), *CAB* and *Capitole* (see Table 5.1) are reported as well, for which Figure 5.3 shows some sample images. Table 6.7 contains for each dataset the largest number (N_e) of connected images after exhaustive pairwise image matching and filtering by the 5-point algorithm and RANSAC, the corresponding number of ROs after filtering N_p, and the number of ROs attributed as correct by each of the five methods.

	N_e	N_p	**Number of correct ROs**				
			PM_RS	**PM_GTC**	**Wang et al. [2019c]**	**Wilson and Snavely [2014]**	**Zach et al. [2010]**
ToH	341	56429	2658	48507	2387	48540	34195
Sta.	156	1733	972	1368	1092	1338	728
Ind.	152	4740	816	4059	1064	3449	3380
Str.	175	5171	1163	3832	1225	4089	4544
Capitole	99	3177	1197	2798	693	1764	2731
CAB	312	10486	2907	4095	2184	4773	6150
PM_RS = The presented method for only detecting outlier ROs due to RS (see Section 3.2.2), as these datasets do not include a critical configuration of VSB;							

Table 6.7: Comparison of the number of ROs from different methods.

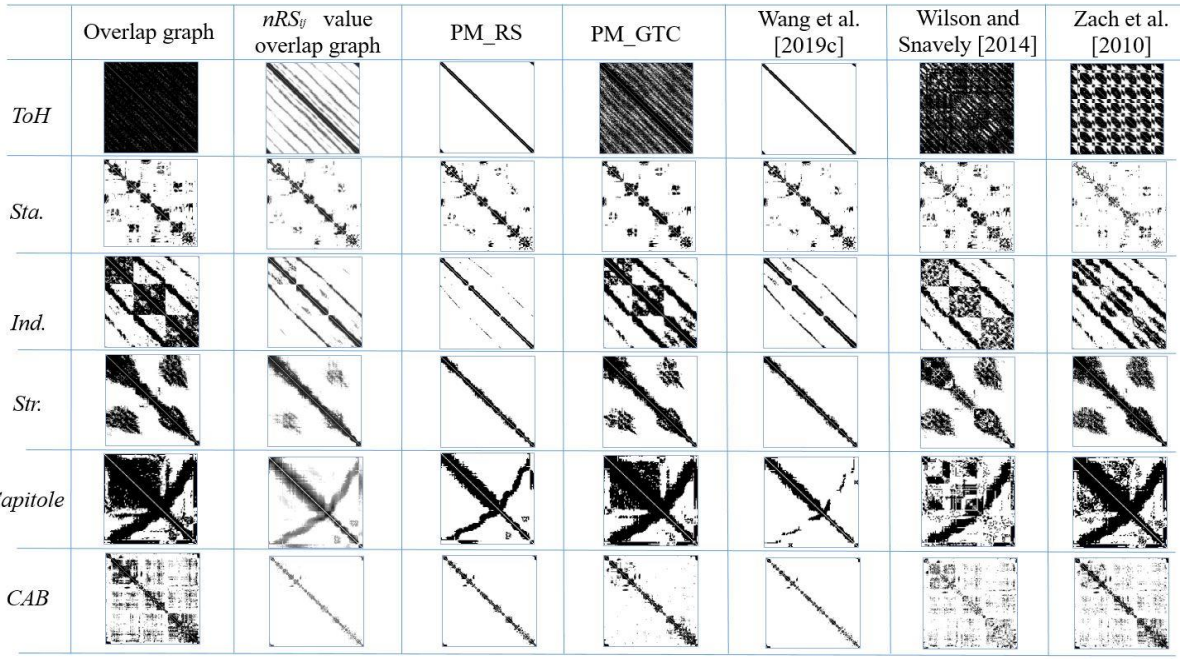


Figure 6.9: Overlap graph of the six public datasets obtained with the different methods. The second column is the overlap graph from the input ROs, black pixels indicate that the corresponding ROs are available. This is also true for the last five columns. The third column is the normalized nRS_{ij} graph.

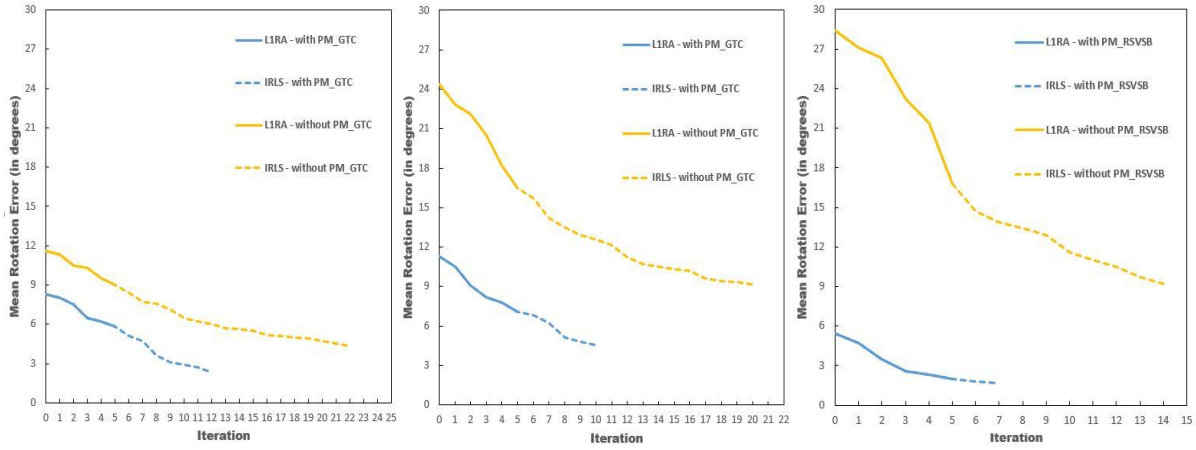
Taking Figure 6.9 and Table 6.7 into consideration, it is found that after the application of the 5-point algorithm each dataset has a very redundant number of ROs, and incorrect ROs survived due to RS. The third column of Figure 6.9 shows the nRS_{ij} graph. Specifically, the nRS_{ij} value is estimated for each RO by transforming the results of Equation (3.2) into the interval [0,1] (by linear normalization) for each dataset individually. The brighter a pixel is, the larger is the corresponding nRS_{ij} and, thus, the higher is the probability that the image pair does not overlap. By investigating the results shown in Figure 6.9, it can be seen that most of the image pairs corresponding to the darker pixels in the third column are kept by PM_RS (the presented method for only detecting RO outliers due to RS) and Wang et al. [2019c], whereas, many pairs related to brighter pixels are still considered as correct using the other three methods. When analysing the numbers in Table 6.7, it can be seen that PM_RS and Wang et al. [2019c] generally eliminate more ROs than the other three methods (this can also be observed by inspecting Figure 6.9: PM_RS and Wang et al. [2019c] result in a clearly thinner overlap graph). Also, the number of ROs selected by Wang et al. [2019c] is always 7 times the corresponding value of N_e (this is a design feature of Wang et al. [2019c]). As a consequence, if fewer correct ROs exist, Wang et al. [2019c] could probably generate incorrect reconstruction results. The self-adapting strategy for parameter selection introduced in PM_RS and the different selection criterion employed in this thesis (see Equation (3.3)) overcome this problem.

Analysing the performance of PM_GTC, PM_RSVSB and PM_RS, PM_GTC shows good capability on the ordered and unordered internet datasets, whereas, it shows very limited success for datasets containing RO outliers due to RS and VSB. Note that the investigations of these public datasets are limited to the direct inspection of ROs and constitute a rather a qualitative evaluation (this is due to the absence of ground truth). The corresponding benefits on the results of global image orientation are explored in the following section and in Section 6.2.3.

Validation of the improved robustness on image orientation

Two further tests are conducted to validate the improved robustness of image orientation results using the refined ROs: *influence on global rotation estimation results* and *influence on the reconstructed object point*.

Influence on global rotation estimation results. Three different datasets from ordered internet (*Lejonet*), unordered internet (*Piazza del Popolo*) and problematic dataset (*B3*) are tested, for which the ground truth rotations of *Lejonet* and *Piazza del Popolo* are provided by the publisher and the results of Wang and Heipke [2020] is adopted as ground truth for *B3*. The performance of the global rotation estimation method [Chatterjee and Govindu, 2013] introduced in Section 4.2 is evaluated by comparing the mean rotation errors of using and not using the proposed robustifying ROs methods. Figure 6.10 shows the convergence behavior of these three datasets, where both *Lejonet* and *Piazza del Popolo* employ the PM_GTC, and *B3* uses the PM_RSVSB.



(a) *Lejonet* (b) *Piazza del Popolo* (c) *B3*

Figure 6.10: Convergence rate with respect to the iteration number.

The yellow curves in Figure 6.10 imply that the original version of Chatterjee and Govindu [2013] can indeed improve the initial rotations. The first five iterations (using *L1RA*) reduce the mean rotation error and provide a good initialization for the subsequent *IRLS* estimator. Nevertheless, after refining the input ROs by the proposed method to increase the robustness of the ROs, several explicit improvements can be seen: First, the initializations obtained by the random MST (Minimum spanning tree) become better. The reason is most probably that the random MST is less contaminated by RO outliers after the robustification method is applied. Especially, for *Piazza del Popolo* and *B3* which are very likely to contain blunders of ROs, the mean rotation errors of the initialization are decreased by a factor between 2 and 5. Second, both the number of iterations necessary for the estimation to converge and the final mean rotation error are highly improved. Without using the proposed method, 22, 20 and 14 iterations are needed for convergence on these three datasets respectively, whereas only 12, 10, and 7 iterations are necessary after employing the proposed method. The final mean rotation errors are reduced by 1.5, 4.6 and 7.5 degrees, respectively. This can be explained by the fact that the improved initialization is efficient for the algorithm *L1RA-IRLS* (cf. **Algorithm 4.3**) in terms of the final solution and the necessary number of iterations. Comparing the runtime listed in Table 6.8, it can be observed that

the overall runtime, and also the average time per iteration, is improved. This behavior is mainly reflected by the procedures of *L1RA* and *IRLS*: the runtime for each corresponding iteration is reduced, which is due to the fact that the number of input ROs becomes smaller after eliminating outlier ROs, which in turn leads to a smaller coefficient matrix when executing *L1RA* and *IRLS* (see **Algorithm 4.1** and **4.2**).

	With proposed method				Without proposed method			
	N_e	L1RA	IRLS	Total	N_e	L1RA	IRLS	Total
Lejonet	368	10.7(2.14)	4.7(0.67)	15.4(1.28)	368	32.5(6.50)	37.3(2.19)	69.8(3.63)
Piazza del Popolo	335	3.3(0.66)	2.1(0.42)	5.4(0.54)	348	6.5(1.30)	22.5(1.50)	29.0(1.45)
B3	342	11.3(2.26)	1.4(0.70)	12.7(1.81)	342	25.7(5.14)	21.6(2.40)	47.3(3.15)

Table 6.8: Runtime in seconds for convergence. The numbers in brackets are the average runtime per iteration.

As a consequence, the proposed ROs robustification methods can indeed positively affect the global rotation estimation method of Chatterjee and Govindu [2013] with respect to both the quality of the final optimal solution and the convergence rate (in particular, the number of iteration and runtime of each iteration are reduced).

Influence on the reconstructed object point. To further demonstrate the advantage of increasing the robustness of ROs, the dataset *facade* provided by Schöps et al. [2017] is tested, for which the ground truth of interior and exterior parameters and the point cloud are provided. To evaluate the number of accurately reconstructed object points, the ground truth interior and exterior orientation parameters are first adopted to calculate the coordinates of all the generated object points, such that the calculated object points are given in the same coordinate system as the ground truth the point cloud. Then, a 3D similarity transformation (see Appendix C) is applied to transform the generated object points to the coordinate system of the ground truth point cloud. The idea of Schöps et al. [2017] is employed to evaluate the estimated results[3]: The authors register the input point cloud to the ground truth point cloud and analyse two criteria, "*accuracy*" and "*completeness*". "*accuracy*" is defined as the fraction of reconstructed points which are located within a distance to the ground truth points smaller than a certain threshold. "*completeness*" is defined as the amount of ground truth points for which the distance to the registered points is below a threshold, where the distance threshold is called "*tolerance*". This thesis only discusses the metric "*accuracy*", because dense matching is not considered to be part of this work, only sparse point clouds are generated, which makes the evaluation of "*completeness*" meaningless. Sparse point clouds are sufficient for the task of image orientation, so that "*accuracy*" is sufficient to carry out the evaluation in this context.

[3] The corresponding project can be accessed via https://github.com/ETH3D/multi-view-evaluation and datasets including images and point clouds can be downloaded via https://www.eth3d.net/datasets

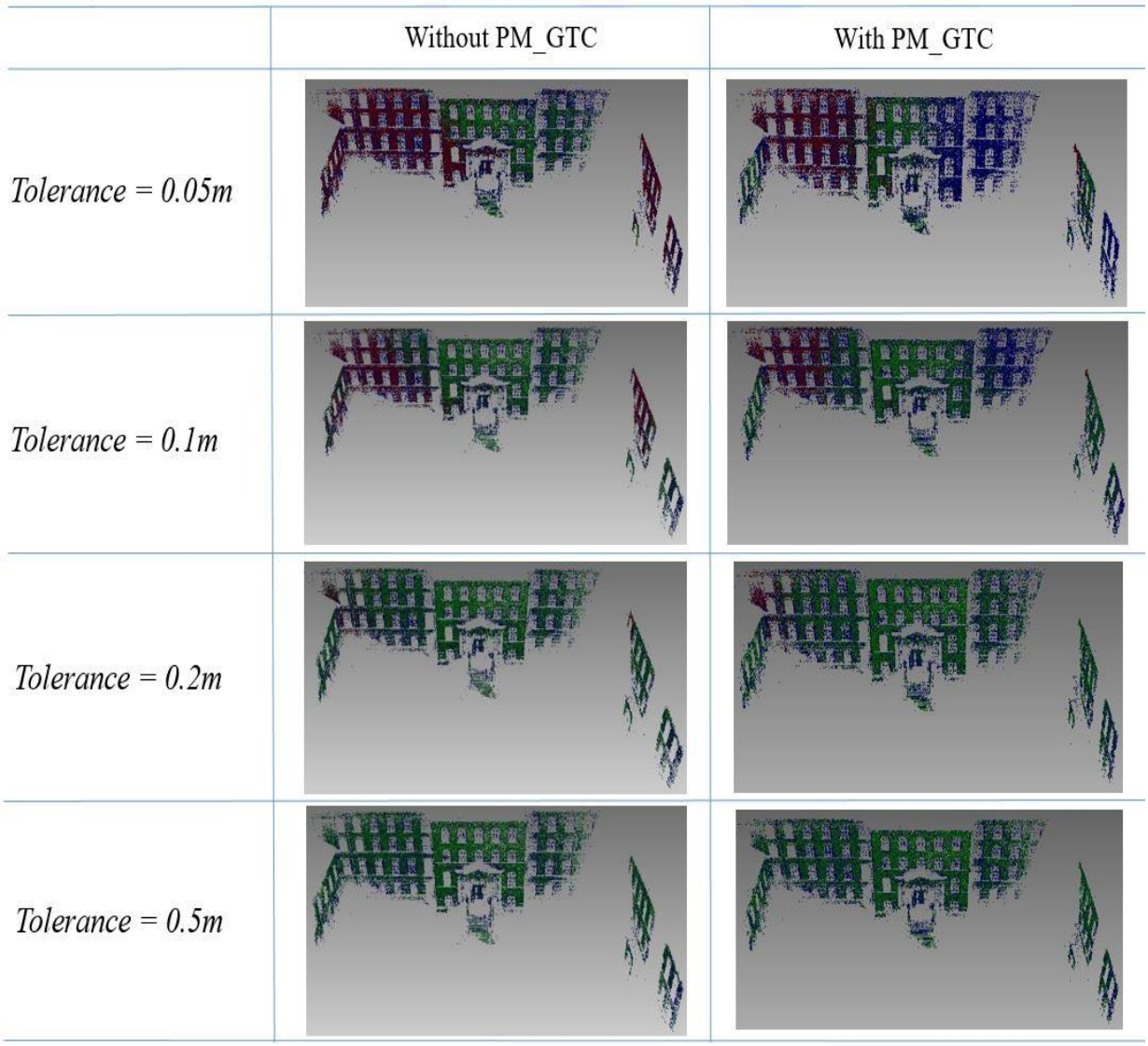


Figure 6.11: Visualization of facade evaluation results. Green points have a position error within the corresponding tolerance value, for red points the position error exceeds the tolerance value, blue points denote that the points cannot be observed by the ground truth point cloud under the corresponding tolerance value by using the method of Schöps et al [2017].

	N_{op}	*Tolerance* (in meters)							
		0.05		0.1		0.2		0.5	
		Accuracy	N_x	*Accuracy*	N_x	*Accuracy*	N_x	*Accuracy*	N_x
Without PM_GTC	90300	38.1%	34407	70.0%	63200	94.9%	85731	99.1%	89487
With PM_GTC	79432	57.1%	45316	81.4%	64672	95.2%	78862	99.3%	79432

Table 6.9: Evaluation on façade using different workflows. N_{op} is the number of object points that are triangulated by the corresponding workflow. N_x is the number of reconstruction points within the distance threshold.

Figure. 6.11 shows the visualized evaluation results of the global image orientation workflows (without PM_GTC and with PM_GTC), where the "*tolerance*" is set to 0.05m, 0.1m, 0.2m and 0.5m. Analysing Figure 6.11 together with Table 6.9, it is found that more object points are triangulated without using PM_GTC because more conjugate points corresponding to the related image pairs can be utilized to generate object points. It can be further seen that using PM_GTC improves the results, as more green points are visible and the obtained accuracy is higher. When

the *tolerance* value is altered to *0.2m* and *0.5m*, as Table 6.9 shows, N_x is larger for Without PM_GTC. This is mainly due to the fact that the total number of reconstructed object points N_{op} is larger, containing many noisy points, from which some object points are kept using a relatively high *tolerance* of *0.2m* or *0.5m*. Although N_x is higher, the "*accuracy*" is lower. Comparing these two workflows, the one enhanced by the proposed ROs robustification procedure performs better. In the case that the tolerance is equal to 0.05m and 0.1m, it provides a larger number of accurate tie points with higher "*accuracy*", while only a smaller number of inaccurate points is kept when the *tolerance* is set to *0.2m* or *0.5m*.

Summary

The performance and efficacy of the proposed methods to increase the robustness of ROs are demonstrated in the previous evaluation experiments. Furthermore, it is also shown that these methods have a positive influence on the image orientation results. Therefore, all the subsequent image orientation experiments are conducted using the relevant ROs robustification methods (see Table 5.1 for the relevant experimental objectives of different datasets).

6.2 Evaluation of global image orientation

The previous sections are dedicated to the evaluation of the proposed pre-processing methods. The general behaviour of finding mutual overlapping image pairs and increasing the robustness of ROs were investigated. In this section, based on the input generated by the proposed pre-processing, the analysis of the global image orientation results is presented. As it has already been stated in Section 5.2, to comprehensively demonstrate that the objective of this thesis can be successfully achieved, various datasets are employed.

For the sake of clarity, the following is structured as: Section 6.2.1 discusses the evaluations on ordered datasets. A set of quasi-oblique aerial images (oblique dataset see Table 5.1) are first tested and the evaluation of image orientation results on 2D image and 3D object space are respectively studied. In addition, a set of ordered datasets from the internet are oriented. In particular, the translation accuracy and the runtime are studied by comparing them with several state-of-the-art global image orientation methods, in particular, two of them are Wang et al. [2019a] and Wang et al. [2021] which were developed by the same author of this thesis, another one is Cui and Tan [2015]. This is followed by an analogous study on unordered internet datasets in Section 6.2.2, in which the translation accuracy and the runtime are again inspected. Finally, the image orientation results on various problematic datasets are shown in Section 6.2.3, for which an extensive qualitative analysis combined with a quantitative evaluation is provided.

6.2.1 Ordered datasets

Two categories of ordered datasets are reported in this section: an oblique dataset and internet datasets.

Oblique dataset

	without bundle adjustment				with bundle adjustment			
	Reference	PM_GTC_IO	(I)	(II)	*Reference*	PM_GTC	(I)	(II)
RMS(x)	-	2.35	2.23	**2.20**	0.14	**0.13**	**0.13**	**0.13**
RMS(y)	-	3.44	3.62	**3.43**	0.15	**0.14**	**0.14**	**0.14**
RMS	-	4.78	4.59	**4.47**	0.20	**0.19**	**0.19**	**0.19**

(I) = Wang et al. [2019a] with L1 norm;
(II) = Wang et al. [2021];
PM_GTC_IO = The proposed global image orientation using the input of PM_GTC.
'-' = the corresponding results are not accessible;
These symbols are used in the following tables and discussions as well.
Reference = Özdemir et al. [2019];

Table 6.10: Precision assessment (in pixels). RMS(x), RMS(y), and RMS are the RMS (root mean square) of reprojection error of the 115 red cross targets (as Figure 5.7 shows) in horizontal and vertical direction and Euclidean residual, respectively.

	without bundle adjustment				with bundle adjustment			
	Reference	PM_GTC_IO	(I)	(II)	*Reference*	PM_GTC	(I)	(II)
CH1	-	**0.91**	0.92	0.94	**0.03**	0.08	0.08	0.09
CH2	-	1.59	1.73	**1.45**	**0.11**	0.15	0.15	0.15
CH3	-	0.85	**0.79**	0.82	**0.03**	0.07	0.06	0.05

Table 6.11: Accuracy assessment (in mm). CH1, CH2, and CH3 are the corresponding check bars shown by Figure 5.7.

	without bundle adjustment			with bundle adjustment		
	PM_GTC_IO	(I)	(II)	PM_GTC_IO	(I)	(II)
RMSE (X)	1.81	1.47	**1.36**	0.56	0.56	**0.56**
RMSE (Y)	5.60	**5.36**	5.48	0.65	**0.65**	**0.65**
RMSE (Z)	1.80	1.96	**1.66**	0.40	0.44	**0.39**
RMSE (O)	1.17	1.17	**1.07**	0.09	0.09	0.09
RMSE (P)	1.34	1.34	**1.02**	0.14	0.14	0.14
RMSE (K)	1.40	1.40	**1.08**	0.25	0.25	0.25

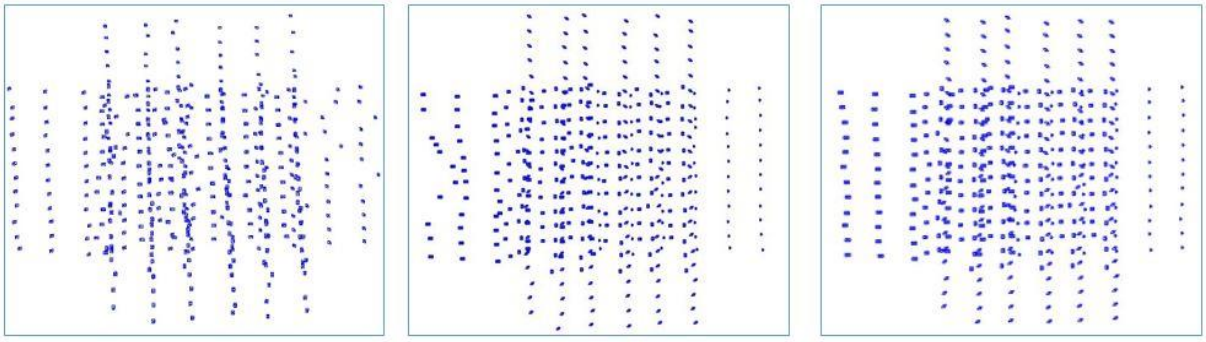
Table 6.12: Relative accuracy assessment. Taking the exterior orientation results of Özdemir et al. [2019] as a reference, RMSE (X), (Y) and (Z) are the root mean square error of translation parameters (mm), RMSE (O), (P) and (K) are the root mean square error of three rotation angles (degrees), O, P and K denote Omega, Phi, Kappa, respectively. The best results are highlighted.

Tables 6.10, 6.11 and 6.12 present the image orientation evaluation criteria on the quasi-oblique aerial images which were explained in Section 5.4.2[1]. As it can be seen by inspecting these three tables, before carrying out bundle adjustment, the root mean square (RMS) value of the reprojection error generated by the proposed method PM_GTC_IO (the proposed global image orientation using the general triplet compatibility check to robustify the input ROs) is around 4.8 pixels (Table 6.10), and the accuracies of the check bars in object space are about 0.8 to 1.8 mm (Table 6.11) which correspond approximately to the GSD ranging from 0.13mm to 0.27mm for

[1] The investigated methods are part of a public image orientation contest (more information is available on http://3dom.fbk.eu/3domcity-task-1-results). The evaluation is only compared to those of other participants. More details of these investigated methods have been introduced in section 2.4.3 and 2.5.

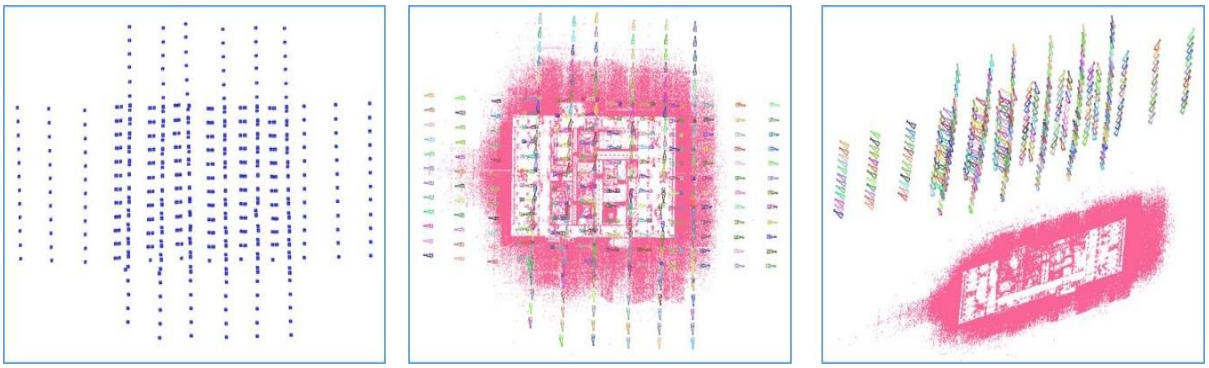
oblique images and 0.12mm for nadir views. The corresponding relative accuracy for exterior orientation parameters before bundle adjustment (Table 6.12) shows the translation error ranges from 1.4mm to 5 mm and the three rotation angle errors are approximately 1.3 degrees. PM_GTC_IO and (I) (indicating the method of Wang et al. [2019a]) have the same rotation error because the same method, Chatterjee and Govindu [2013], is applied to provide global rotation solution. (II) (indicating the method of Wang et al. [2021]) is slightly better than the other two methods on these criteria, because (II) benefits from the fact that only a subset of the best triplets (just enough to connect all images) is considered, and the corresponding ROs within the selected triplets used as inputs are more robust with respect to the triplet compatibility. A qualitative comparison can be seen in Figure 6.12 in which the exposure positions of images are illustrated. All methods generally produce initial values for bundle adjustment that are approximately coincident with the reference (see Figure 6.13 (a)). However, the exposure positions of (II) are visually closer to the reference which is consistent with the results shown in Tables 6.10-6.12.

After bundle adjustment, almost the same precision is achieved by all three methods. This in turn implies that all of these methods can be employed as a tool for providing a reliable initialization for the final bundle adjustment and almost the same optimal solution is achieved. Figure 6.13 (b) and (c) are the visualization of final reconstruction results from two different perspectives (after bundle adjustment) using PM_GTC_IO.



(a) (b) (c)

Figure 6.12: Visualization of exposure positions before bundle adjustment. (a) is the result of PM_GTC_IO. (b) and (c) are the results of (I) and (II), respectively.



(a) (b) (c)

Figure 6.13: Visualization of the result after bundle adjustment. (a) is the exposure positions from the provided reference. (b) and (c) are the reconstruction results of PM_GTC_IO observed from two various perspectives.

Internet datasets

Name	Without BA			With BA			
	PM_GTC_IO	(I)	(II)	Incre.	PM_GTC_IO	(I)	(II)
Porta	1.055	1.236	**0.743**	**0.063**	0.075	0.107	0.074
CouCha	1.103	**1.056**	1.342	**0.095**	0.106	0.125	0.105
Gbg	**0.461**	0.714	0.481	**0.066**	0.081	0.088	0.069
Pumpkin	**0.861**	1.123	1.313	**0.061**	0.077	0.079	0.079
SriMar	0.536	0.753	**0.521**	**0.025**	0.060	0.073	0.059
Fapalace	**0.156**	0.352	0.291	**0.036**	0.046	0.056	0.047
Ystad	0.789	0.951	**0.566**	**0.051**	0.077	0.093	0.078
Buddah	1.667	1.891	**1.091**	0.781	0.767	0.779	**0.740**
Kingscollege	**0.863**	1.135	0.991	0.185	**0.121**	0.147	0.126
Lejonet	2.334	3.341	**1.913**	0.344	0.331	0.347	**0.328**
UWO	**0.761**	1.059	0.934	**0.079**	0.082	0.088	0.080
Orebro	**0.777**	1.234	1.189	0.051	**0.045**	0.050	**0.045**
Spilled	**1.120**	1.534	1.743	0.108	0.105	0.134	**0.103**
Ahus	**3.256**	3.753	3.941	**0.291**	0.316	0.345	0.307
SanMarc	1.112	1.591	**0.593**	**0.077**	0.092	0.093	0.092
Duomo	**1.056**	1.444	1.346	**0.120**	0.120	0.151	0.145
Incre. = The incremental image orientation method of Wang et al. [2018]. This abbreviation is also used in the following tables.							

Table 6.13: Translation evaluation on ordered internet datasets. The mean translation error of the various methods is listed, which is given in a random unit defined up to an unknown similarity transform[2]. The best results in each row are highlighted.

In this section, the abovementioned global image orientation methods are compared using the ordered internet datasets (see Table 5.1) and the translation errors before and after applying the bundle adjustment are discussed. Furthermore, to demonstrate the advantage of the global image orientation method's time efficiency, an incremental image orientation method (Wang et al. [2018]) is included for comparison, obviously, it is only feasible to analyze the results after bundle adjustment, because images are sequentially oriented and results without any refinement of bundle adjustment, in principle, cannot be obtained. Detailed quantitative results are shown in Table 6.13. From the results without bundle adjustment, all three methods initialize the exterior orientations with small differences which can be attributed to the relatively good input ROs, as all these datasets were captured by professionals (Olsson and Enqvist [2011) using a camera of relatively good quality (the initial intrinsic parameter from EXIF file is relatively accurate) and a suitable overlap degree among images was taken into account during image capture. Even so, the proposed PM_GTC_IO method is generally better than (II) in estimating translations (only a few datasets are worse, and with just a very small margin), while (I) is the worst. In 9 out of 16 datasets, the best results are obtained by PM_GTC_IO, however, (II) is superior on the other 6 datasets, (I) delivers the best solution only for one dataset. This can be explained by the fact that the proposed method applies the general ROs robustification method considering both rotation and translation triplet capability, and (II) only employs a subset of some best triplets, however, only rotation triplet

[2] The datasets together with the references are available at http://www.maths.lth.se/matematiklth/personal/calle/dataset/dataset.html. The reconstruction results provided by the authors (Olsson and Enqvist [2011]) are considered as references. These results are all provided with random units which are related to a specific coordinate system up to an unknown similarity transformation. Rescaling, rotating or translating the reconstruction does not change the reprojections. Therefore, it makes sense to measure the quality of the orientation result by the distance between the corresponding solution and the reference.

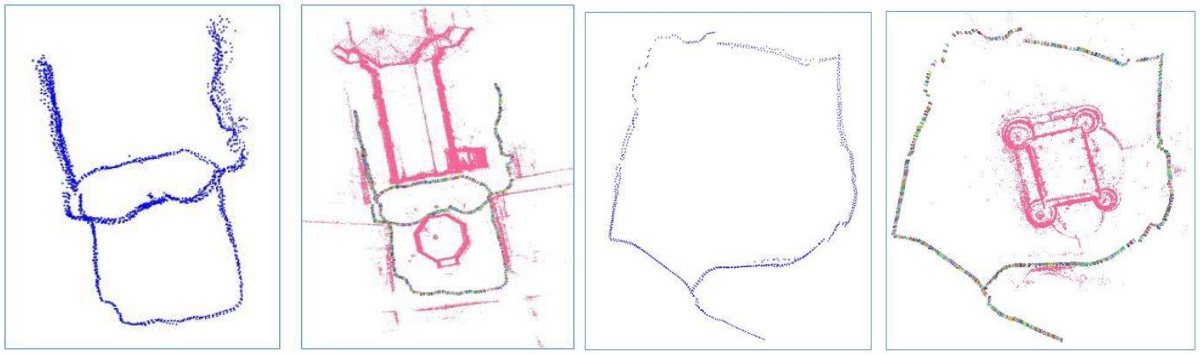
capability is utilized in (I) and many tie points are selected to connect all images for solving translations, outliers in these selected tie points can negatively affect the solution. After the refinement by bundle adjustment, as Table 6.13 illustrates, all translation accuracies from the various methods are improved. In general, superior results can be observed for the incremental method, because incremental methods are supposed to be more robust and outliers are recursively handled when iteratively adding new images and running bundle adjustment. In addition, similar to the corresponding results on the oblique dataset, there are no clear differences among these three global image orientation methods. This indicates that they are all able to provide a reliable initialization which makes the final bundle adjustment converge to a similar optimal solution. The qualitative visualizations of the ordered internet datasets are shown in Figure 6.14, in which the figure showing the blue trajectory denotes the exposure positions computed by the proposed method PM_GTC_IO without bundle adjustment, and the figure with red dots and colorful triangles is the corresponding refined reconstruction result after bundle adjustment.

Name	Incre.		PM_GTC_IO				(I)				(II)			
	N_r	T_Σ	N_r	T_o	T_{BA}	T_Σ	N_r	T_o	T_{BA}	T_Σ	N_r	T_o	T_{BA}	T_Σ
Porta	141	396	141	37	99	136	141	187	112	299	141	20	93	**118**
CouCha	176	535	176	25	175	**200**	176	63	177	240	176	25	185	210
Gbg	179	636	179	78	90	168	179	67	84	151	179	25	96	**121**
Pumpkin	209	753	209	199	124	323	209	178	158	336	209	44	112	**156**
SriMar	222	1622	222	38	154	**192**	222	79	144	223	222	45	163	208
Fapalace	281	1736	275	138	188	326	281	346	171	517	281	52	201	**253**
Ystad	290	2133	290	179	254	433	290	321	193	514	290	59	233	**292**
Buddah	322	1985	322	225	200	425	322	478	256	734	322	70	182	**252**
Kingscollege	355	3145	355	207	170	377	355	433	195	628	355	79	156	**235**
Lejonet	368	3466	368	256	169	427	368	567	212	779	368	85	169	**254**
UWO	692	9074	688	561	395	956	692	954	432	1386	692	134	369	**503**
Orebro	763	13423	763	726	351	1077	763	1159	444	1603	763	188	333	**521**
Spilled	781	14891	780	742	439	1181	781	1334	476	1810	781	172	414	**586**
Ahus	805	15116	801	736	444	1180	805	1064	774	1838	805	246	399	**645**
SanMarc	1499	19921	1466	840	523	1363	1479	1442	601	2043	1499	297	533	**830**
Duomo	1805	26314	1793	1000	783	1783	1796	1857	813	2670	1805	376	763	**1139**

Table 6.14: Runtime in seconds for ordered internet datasets by using different methods. N_r is the number of orientated images. T_o is the time for image orientation, T_{BA} denotes the runtime for bundle adjustment. T_Σ indicates the total runtime. The fastest approach in each row is highlighted.

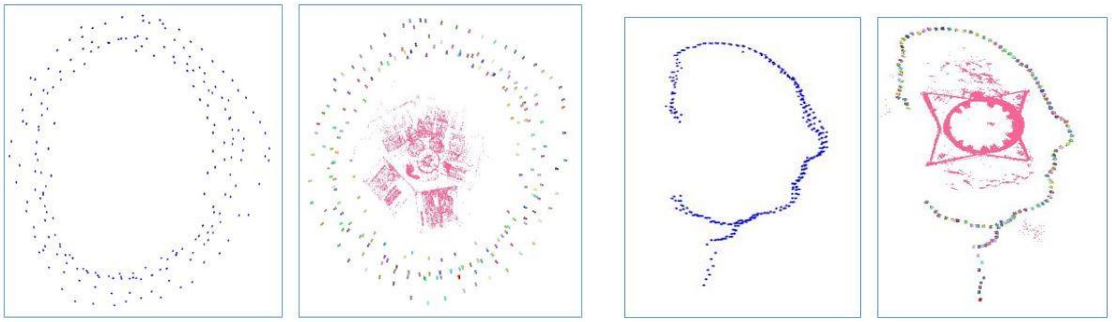
Table 6.14 lists the runtime for completing the image orientation task on ordered internet datasets. Again, all four methods were tested on the same hardware (a quad-core processor (3.2 GHz Inter (R) Core (TM)i5-6500, 32G memory) with eight threads available in total). Although the incremental method gives the best image orientation results, it is the slowest by a factor of around 2 to 20 (depending on the size of datasets). Carrying out investigations among the global methods, their performance with respect to the runtime is within the same order of magnitude. However, some notable findings can be summarized: first, (II) is generally the most time-efficient solution and (I) is the slowest one, which is mainly due to the runtime required to solve the initial image orientation (see the related T_o column of Table 6.14). (II) solves the initialization based on a subset of minimal best connected triplets and the exterior orientation parameters (rotations and translations) are calculated synchronously. On the other hand, (I) solves a large linear equation

system which estimates every image translation and the necessary selected tie points simultaneously. Furthermore, the L_1 norm is applied for a robust estimation which is not as fast as the L_2 norm. In this thesis, all remaining ROs are used to solve the unknown translations, it is thus inferior to (II), on the other hand, no tie point coordinates need to be solved and the L_2 norm is applied as well, therefore the proposed PM_GTC_IO is faster than (I). Second, the number of recovered images varies for different methods: while the incremental method and (II) typically solve all the images, PM_GTC_IO and (I) orient slightly less images. This is because some images that are not well connected to the photogrammetric block are excluded by the triplet compatibility check. Specifically, PM_GTC_IO excludes more images than (I) as both rotation and translation compatibility constraints are considered by PM_GTC_IO and (I) only employs the rotation compatibility constraint. For the incremental method, the ROs after epipolar geometric validation are directly inserted to image orientation (see relevant numbers in Table 6.2); as for (II), the completeness of the photogrammetric block is inherently guaranteed when dealing with triplet selections.



Duomo *Orebro*

SanMarc *Ahus*



Pumpkin *Lejonet*

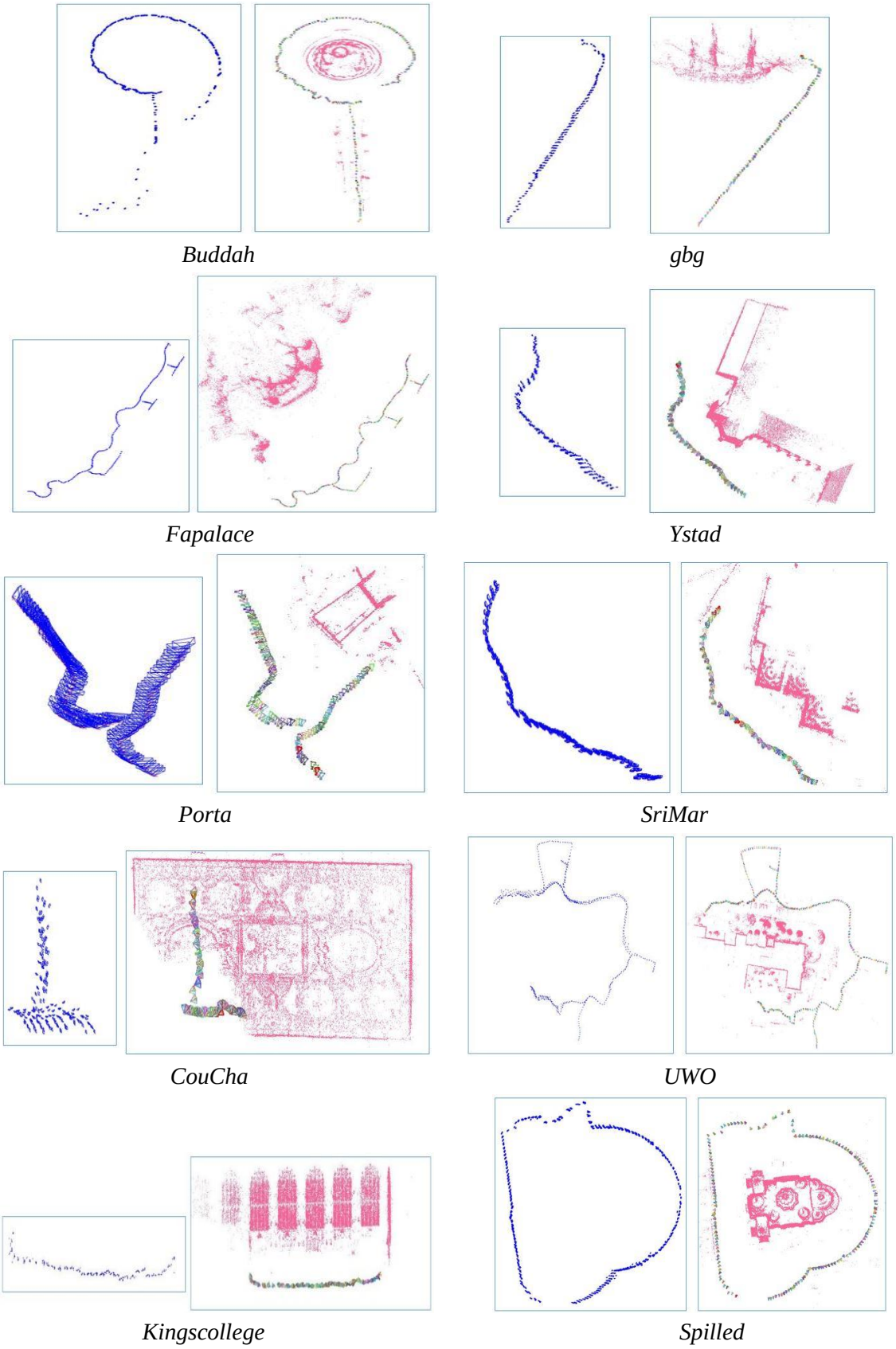
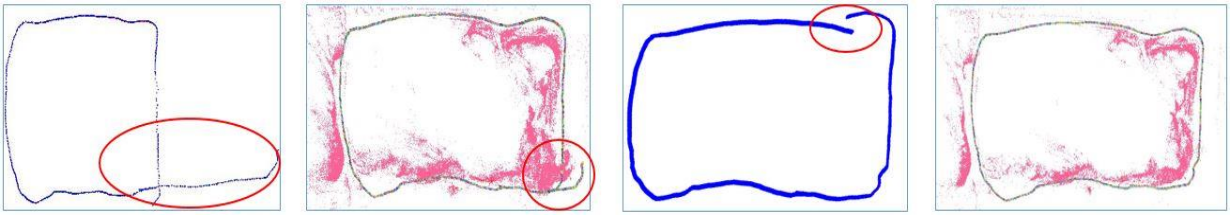


Figure 6.14: Visualization of the results for ordered internet datasets. The blue trajectory in the left part per dataset denotes the exposure positions computed by the proposed method PM_GTC_IO without bundle adjustment, and the figure with red dots and colorful triangles is the corresponding refined reconstruction result after bundle adjustment

Summary

Based on the described results of ordered datasets, it can be observed that the incremental method normally performs best, but it costs much more time to carry out the image orientation task. Nevertheless, a fairly similar accuracy can be achieved using global methods (after bundle adjustment) which is 2 to 20 times faster. Comparing the proposed PM_GTC_IO and (II), it is found that (II) is generally faster than PM_GTC_IO without large discrepancies with respect to the translation error ((II) is also better on some datasets), therefore (II) seems to diminish the significance of this thesis. Yet, the proposed method shows good results on a dataset which (II) fails to deal with. This dataset, namely *campus* (provided by Cui and Tan [2015]), is made up of 1040 images which form a very large closed loop. From Figure 6.15(a), it can be seen that (II) produces a very distinct drift, resulting from the fact that all image pose parameters are estimated by hierarchically traversing the selected connected triplets. The error from the similarity transformation between two connected triplets is recursively accumulated during the traversing procedure, thus, the estimated initial image pose parameters are too far away from the global optimum. Consequently, even after bundle adjustment the drift is only reduced but not eliminated. In contrast, the disconnection of the initial loop in PM_GTC_IO is very small. Moreover, the reconstruction visually result shows a closed loop after bundle adjustment. This is mainly due to the fact that all eligible relative orientations are applied for solving exterior orientation parameters and errors are evenly distributed on every estimated image pose, which also means that the loop closure constraint is implicitly considered by this manner. This inherent characteristic can be more explicitly demonstrated on the following unordered datasets, as loops would be inadvertently formed if images were taken in an arbitrary way (any two images are possible to overlap).



(a) (b)

Figure 6.15: Visualization of image orientation results of campus. (a) is the result of (II). (b) is the result of the proposed PM_GTC_IO, the blue trajectories denote the exposure positions before bundle adjustment. The red ellipses denote the visual drifts.

6.2.2 Unordered datasets

Analogous to ordered internet datasets, a similar numerical evaluation on unordered internet datasets is carried out and shown in Table 6.15. The reference results[3] are generated by using Bundler [Snavely et al., 2006], as this strategy is widely used by many other global SfM methods. The evaluation of (III) [Cui and Tan, 2015], which also employs information of relative translations and tie points, is directly cited from the corresponding paper, while the others are self-implemented on the same machine. Before bundle adjustment, the proposed PM_GTC_IO

[3] See more information at http://www.cs.cornell.edu/projects/1dsfm/

performs best on most (10 of 14) datasets, (II) and (III) are better on just two and three of them, respectively. Investigating the results of (I), it can be seen that this method is the worst. The reason for this is basically identical to the one on the previously discussed ordered internet datasets, but with a few particular factors relating to unordered internet datasets: the selected tie points have a higher chance (compared to ordered internet datasets) to be contaminated by spurious matches and the initial stereo pair (which can have a significant influence on the quality of the orientation result) might not be well selected due to the fact that complex stereo configurations exist in unordered internet datasets. On the other hand, the impact of the implicitly considered loop closure constraints is revealed on these unordered datasets, which allows the proposed method to typically outperform (II). Note also, that (I) and (II) failed to deal with *Piccadily* because of the limitation of the used machine in terms of main memory. The proposed method can, however, bypass this limitation. After the bundle adjustment, it is again found that the incremental method generates the best translations on most datasets. However, the three self-implemented global methods obtain a similar performance which in turn implies that the final bundle adjustment finds comparable optimal solutions using the initialization from these three global methods. Lastly, via inspecting (III), before bundle adjustment the accuracy differences between (III) and the proposed method are small, whereas, after bundle adjustment PM_GTC_IO is better than (III) and some results of (III) become even worse. The reason is probably that the presented robust bundle adjustment used here is more efficient than the one employed by (III). Visualizations of the unordered internet dataset reconstruction results are shown in Figure 6.16.

Name	Without BA				With BA				
	PM_GTC_IO	(I)	(II)	(III)	Incre.	PM_GTC_IO	(I)	(II)	(III)
Ellis_Island	**1.9**	4.6	10.7	5.5	**0.8**	1.1	1.3	2.1	4.2
Piazza del Popolo	**2.1**	5.4	7.4	2.7	2.1	**1.8**	2.5	2.6	2.5
NYC Library	**1.6**	2.0	2.6	1.9	**0.9**	1.4	1.3	**1.0**	1.6
Metropolis	**7.4**	9.8	11.6	10.6	2.2	**2.1**	2.6	2.2	16.6
York Minster	**3.2**	5.9	6.1	5.7	2.1	2.9	3.6	**1.9**	14.2
Montreal ND	1.1	1.9	1.5	**0.7**	**0.5**	0.9	0.9	0.6	1.1
Tower of London	**9.6**	12.4	11.7	11.2	**1.6**	2.0	4.0	3.1	12.5
Notre Dame	1.7	2.9	1.3	**0.6**	0.7	1.4	1.6	0.3	1.0
Alamo	**0.7**	1.4	2.6	2.0	**0.3**	0.5	0.5	0.6	3.1
Gendarmenmark	**20.3**	25.1	25.5	27.7	**9.3**	12.4	12.9	12.4	27.3
Vienna Cathedral	**3.9**	7.6	6.2	5.9	**2.6**	3.1	3.9	4.2	4.9
Union Square	**6.4**	7.1	**6.4**	12.7	**3.9**	4.4	4.4	4.3	11.7
Roman Forum	10.2	11.3	**9.3**	9.4	6.1	**2.9**	5.8	5.5	10.1
Piccadily	4.1	-	-	**2.5**	**1.9**	2.1	-	-	2.2

(III) = Cui and Tan [2015], this item is used in the following tables and contents.
"-" = The corresponding results are not available.

Table 6.15: Translation evaluation on unordered internet datasets. The mean translation errors (in meters) of various methods are provided. The best results in each row are highlighted.

Table 6.16 provides the individual runtimes of the unordered internet datasets evaluated. Note that the runtime of Cui and Tan [2015] is in fact not directly comparable because of the different hardware that was used to obtain these numbers. However, it is still interesting to see how large the gap is between the proposed method and the state-of-the-art method[4] regarding time efficiency.

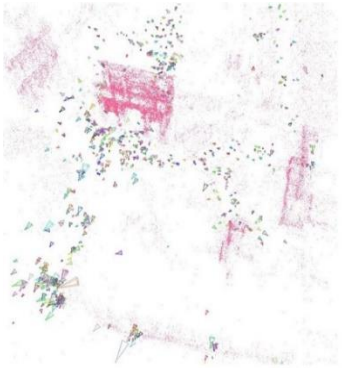
[4] Cui and Tan [2015] ran their experiments on a machine with two 2.3 GHz Intel Xeon E5-2650 processor with 16 threads enabled.

From Table 6.16, (III) performs faster on 8 of 14 datasets. Most of them benefit from the lower time the bundle adjustment needed (it is worth to mention that four out of these eight datasets have worse translation accuracy after bundle adjustment than before); PM_GTC_IO and (II) run fastest on three other datasets each. To allow for a fair comparison, only the self-implemented methods are studied in more detail, as they are run on the same machine. In general, all global methods are always faster than the incremental method, normally around 10-20 times depending on the specific dataset. The reduced runtime mainly results from the computationally expensive intermediate bundle adjustment that is carried out by the incremental method. PM_GTC_IO and (II) are superior to (I), and (II) is faster than PM_GTC_IO on 8 of 14 unordered datasets. In addition, approximately the same number of images are solved, but (II) solves slightly more. These observations are basically consistent with the ones on the results of ordered internet datasets shown in Table 6.14.

Name	Incre.		PM_GTC_IO				(I)				(II)				(III)	
	N_r	T_Σ	N_r	T_o	T_{BA}	T_Σ	N_r	T_o	T_{BA}	T_Σ	N_r	T_o	T_{BA}	T_Σ	T_{BA}	T_Σ
Ellis_Island	219	1556	221	30	85	**115**	219	73	101	174	223	41	95	136	169	208
Piazza del Popolo	300	3116	304	24	188	212	299	91	189	280	309	64	139	203	147	**194**
NYC Library	293	3046	305	20	184	204	310	86	176	262	311	62	136	**198**	171	213
Metropolis	284	2314	296	15	242	257	290	81	227	308	306	63	185	248	25	**60**
York Minster	386	4751	381	35	538	573	377	142	574	716	388	89	405	**494**	611	663
Montreal ND	431	6136	427	75	493	568	427	182	469	651	433	98	396	**494**	613	648
Tower of London	407	5966	404	36	342	**360**	414	102	366	468	421	81	350	431	503	563
Notre Dame	529	9364	519	230	884	1114	520	277	871	1148	539	134	856	990	461	**552**
Alamo	522	7801	532	124	253	**359**	526	244	321	566	535	120	244	364	481	578
Gendarmenmark	488	3321	472	30	344	374	477	92	374	466	477	103	393	496	131	**214**
Vienna Cathedral	722	14331	736	169	884	1053	713	394	834	1228	736	169	754	903	440	**582**
Union Square	691	7421	704	124	233	357	689	166	338	504	726	173	282	455	47	**92**
Raman Forum	934	19752	973	160	1294	1454	964	311	1313	1624	994	268	984	1252	339	**491**
Piccadily	1913	27064	1856	1226	1029	2255	-	-	-	-	-	-	-	-	1053	**1480**

Table 6.16: Runtime in seconds for unordered internet datasets. N_r is the number of orientated images. (III) is directly cited from the paper Cui and Tan [2015]. T_o is the time for image orientation, T_{BA} denotes the runtime for bundle adjustment. T_Σ indicates the total runtime. The fastest approach in each row is highlighted.

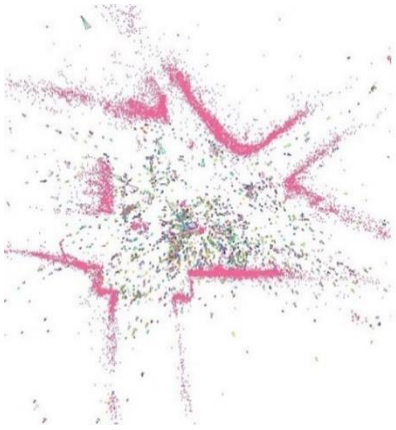
Montreal Notre Dame *Notre Dame* *Alamo*



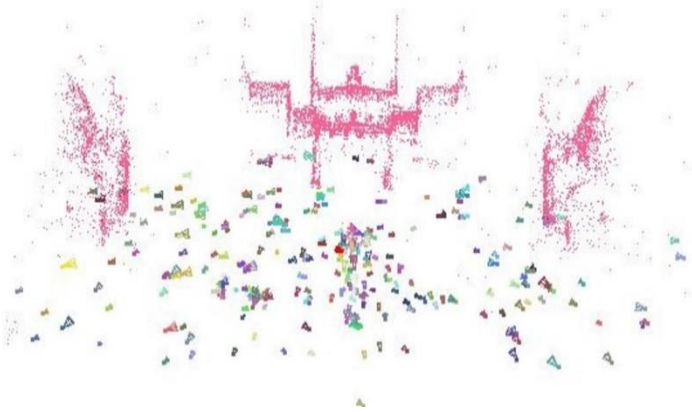
Roman Forum

Union Square

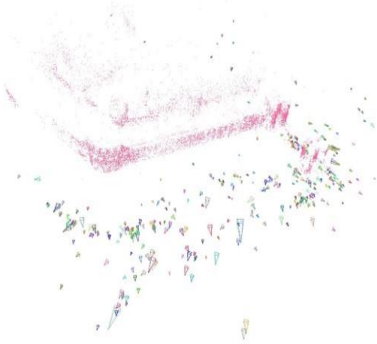
NYC Library



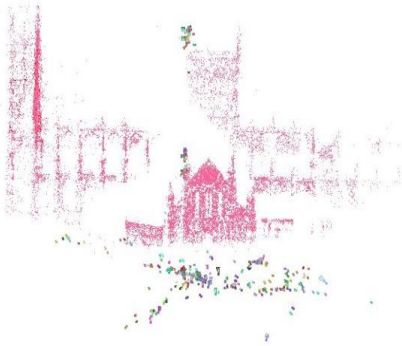
Piccadily



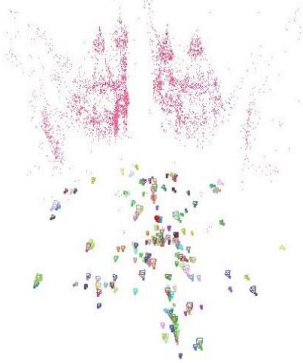
Gendarmenmarkt



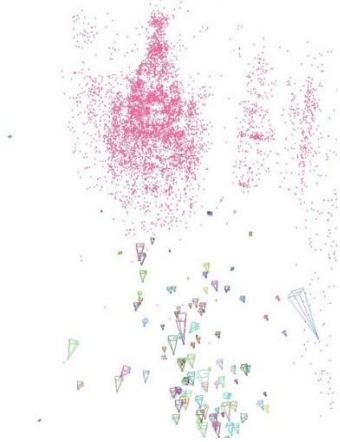
Tower of London



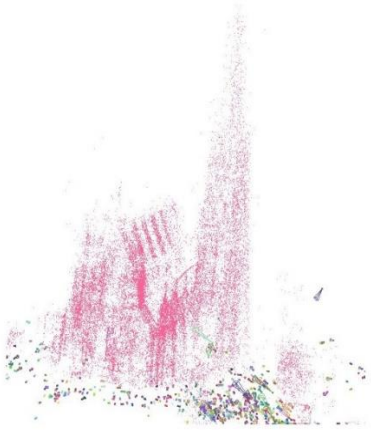
York Minster



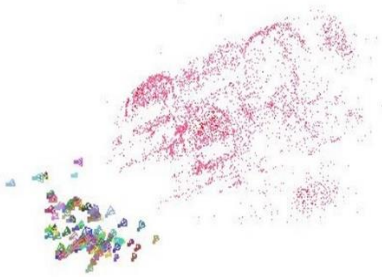
Piazza del Popolo



Metropolis



Vienna Cathedral



Ellis Island

Figure 6.16: Visualization of the results for nine of the internet datasets.

6.2.3 Problematic datasets

Benchmark with RS and VSB ROs

To investigate how RS and VSB ROs affect the proposed global image orientation method and to demonstrate that the elimination of wrong ROs is beneficial to improve the reconstruction results, based on the benchmarks with RS and VSB ROs, validation experiments are conducted using five pipelines with different sets of ROs as input (the ROs generated from different pipelines are fed into the proposed global method for image orientation): the manually generated ground truths ROs, the presented RS and VSB RO elimination method together with BPVD RO identification (these two are indicated by "GT_IO" and "PM_RSVSB_IO", respectively, in the following content and figures), no ROs elimination, only RS elimination, and only VSB elimination with BPVD RO identification, which are denoted as "PM_Noclean_IO", "PM_RS_IO" and "PM_VSBBPVD_IO", respectively. Figure 6.17 visualizes the reconstruction results (Note that ground truth reconstruction results are not available.). Images from google maps show the individual scene and the footprints of these buildings are highlighted by green lines. Not surprisingly, using the ground truth ROs, the proposed GT_IO can produce reconstruction results which are consistent with the footprint images (this again demonstrates the effectiveness of the specific proposed global image orientation method). Comparing "PM_RSVSB_IO" to "GT_IO", it can be found that there are no visual drifts between them which implies that the ROs selected by the proposed methods are mostly correct. Investigating the results of "PM_RSVSB_IO" and of the other three experiments further, it becomes obvious that only the proposed method yields correct results, as artefacts were generated by the other three methods (shown in the ellipses). Reminding that, although BPVD ROs identification was conducted on these three benchmarks, there exist in fact no BPVD ROs in these three benchmarks and none of the detected VSB ROs were identified as BPVD ROs. It can be concluded that both, RS and VSB ROs have a negative effect on the presented global image orientation method. Thus, it is necessary to eliminate both types of errors.

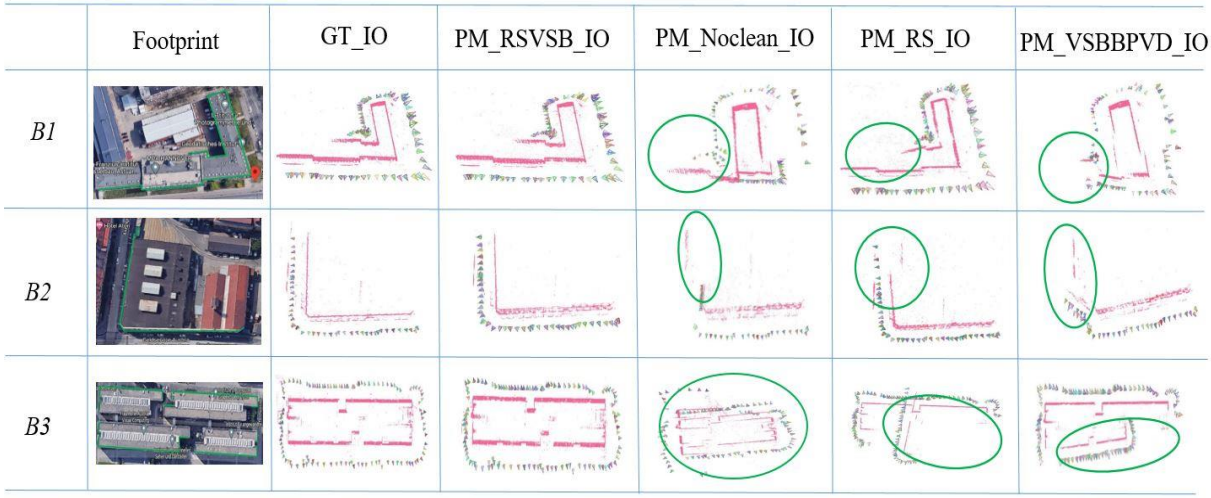
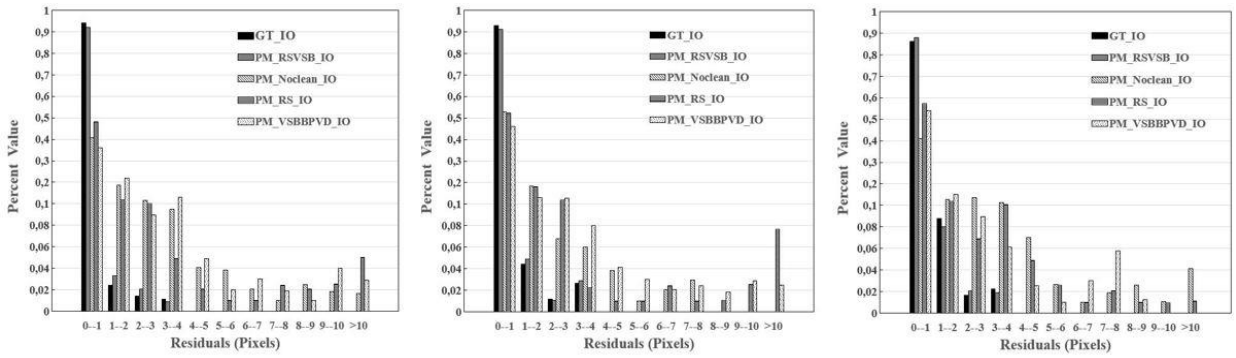


Figure 6.17: Visualization of reconstruction results from the five different pipelines. The footprint images were downloaded from google maps (green lines highlight the footprint of the buildings). Ellipses denote visual artefacts.

The visualizations of the reconstruction results in Figure 6.17 are only shown for a qualitative comparison, a related numerical analysis is further provided. Figure 6.18 shows the reprojection error distribution of the three benchmarks from the five different pipelines. According to the comparison, there are only very small discrepancies between GT_IO and PM_RSVSB_IO which can also be indirectly implied from Figure 6.17. Nearly 90% of the reprojection errors are within one pixel for GT_IO and PM_RSVSB_IO, and the maximum reprojection errors lie between 3 and 4 pixels, whereas for the other three pipelines, there are much fewer reprojection errors smaller than one pixel and significantly more between 1 pixel and 4 pixels. The maximum reprojection errors are larger than 10 pixels in most cases. Therefore, eliminating incorrect ROs by the suggested method does significantly improve the robustness of the presented global image orientation method.



(a) *B1* (b) *B2* (c) *B3*

Figure 6.18. Reprojection error distribution of five different pipelines on the three benchmarks.

Figure 6.19 shows a visualization of the reconstruction results obtained by using the ROs considered to be correct by different methods of the literatures, for which the results of detected overlap graphs are shown in Figure 6.8 in Section 6.1.2, also, the presented global image orientation method is applied; Artefacts, depicted as ellipses, are again visible in the results of Zach et al. [2010], Wilson and Snavely [2014] and PM_GTC_IO. It can be concluded that the proposed RO elimination method generates the best results.

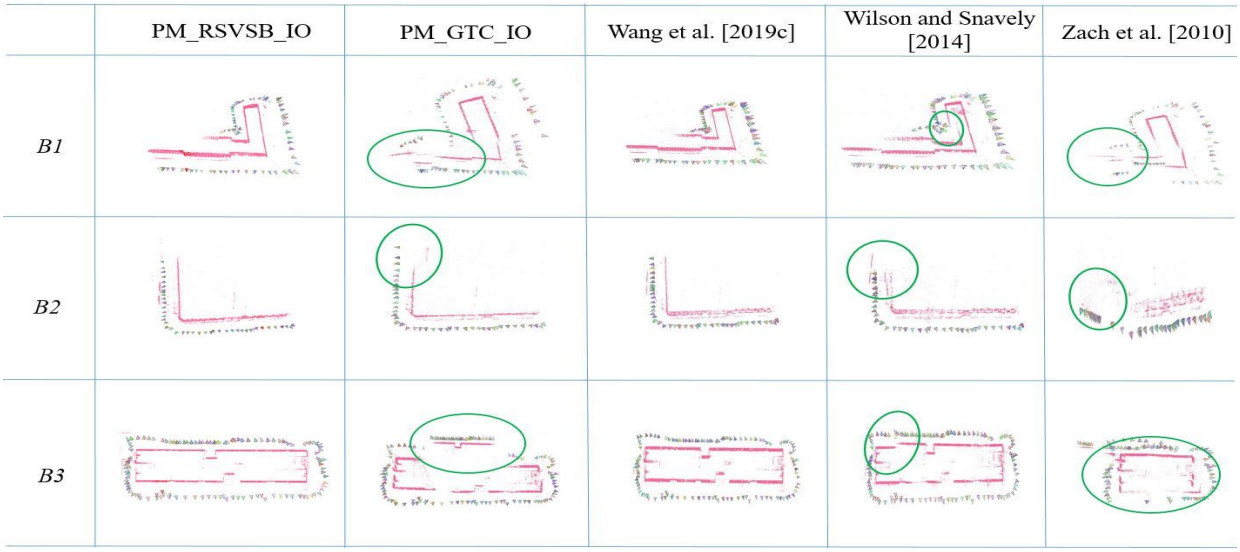
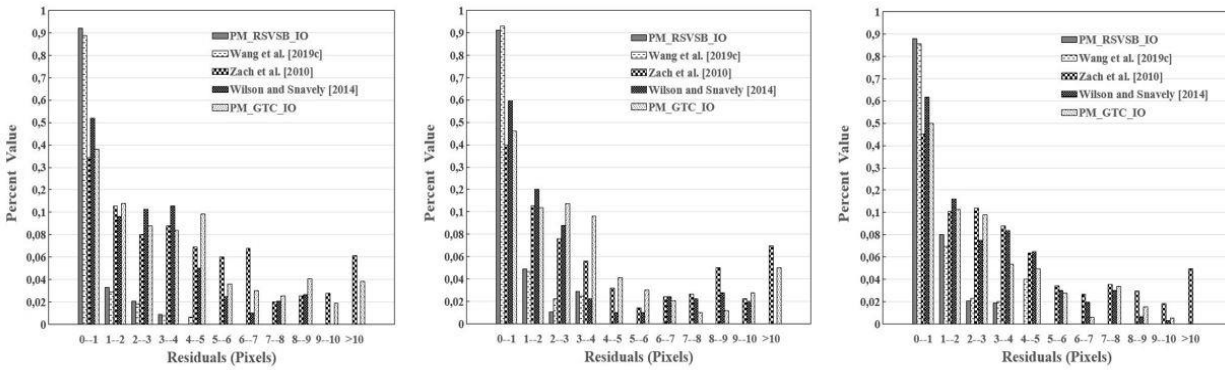


Figure 6.19: Visualization of reconstruction results on three benchmarks from different methods.

Figure 6.20 again shows the distribution of the reprojection errors obtained with the different methods on these three benchmarks. PM_RSVSB_IO and Wang et al. [2019c] yield very similar results, approximately 90% of the errors are again below one pixel and the maximum residuals are smaller than 5 pixels. The reprojection error distributions of Zach et al. [2010], Wilson and Snavely [2014] and PM_GTC_IO are again wider, reprojection errors larger than 8 pixels exist in all benchmarks, and the proportions of reprojection errors smaller than one pixel are much lower than those of PM_RSVSB_IO and Wang et al. [2019c].



(a) *B1* (b) *B2* (c) *B3*

Figure 6.20: Reprojection error distribution of different methods on the three benchmarks.

Comparing the results of PM_RSVSB_IO and Wang et al. [2019c] shown in Figures 6.19 and 6.20, it can be concluded that the performances of these two methods on these three benchmarks is very similar. It is interesting to note, however, that Wang et al. [2019c] selected a set of individual free parameters for each benchmark dataset to obtain the reported results (as mentioned, these individual parameters are used for the results reported here). These free parameters are rather difficult to determine in advance, and the results of Wang et al. [2019c] are sensitive to their choice. In contrast, PM_RSVSB_IO does not have such difficulty and adopts a self-adapting strategy for parameter selection when dealing with RO outliers due to RS and VSB.

Public datasets with high degree of RS

The public datasets with a high degree of RS are also assessed. Again no ground truth for the ROs is provided for these datasets to independently validate the quality of the detected correct ROs. Similar to the evaluation of the previous sections, different sets of ROs are inserted into the suggested global image orientation pipeline. First, a comparison of PM_RS_IO, Zach et al. [2010], Wilson and Snavely [2014] and PM_GTC_IO is discussed for *ToH*, *Sta*., *Ind*. and *Str*. Visualizations of the reconstruction results are shown in Figure 6.21, the ellipses denote artefacts. Th reconstruction results of Wang et al. [2019c] are visually identical with those of PM_RS_IO. For *ToH*, only a part of the temple is reconstructed by Zach et al. [2010], Wilson and Snavely [2014] and PM_GTC_IO. Probably due to the RS ROs in the *Str*. dataset, which results in an overlap graph with a pair of wings as shown in Figure 6.9, Zach et al. [2010], Wilson and Snavely [2014] and PM_GTC_IO all generated a folded reconstruction. As for *Ind*., many images are incorrectly oriented by Zach et al. [2010], Wilson and Snavely [2014] and PM_GTC_IO, so that these three methods again produce a folded reconstruction result. The reconstruction result of *Sta*. by Zach et al. [2010] does not keep a consistent block, and it has the lowest number of ROs (728);

the circular stadium is also not closed by Wilson and Snavely [2014] or by PM_GTC_IO. In contrast to the other methods, PM_RS_IO does not show any visual artefact on these four datasets. This proves the capability of the proposed method to detect RS ROs and to consequently deliver correct reconstruction results.

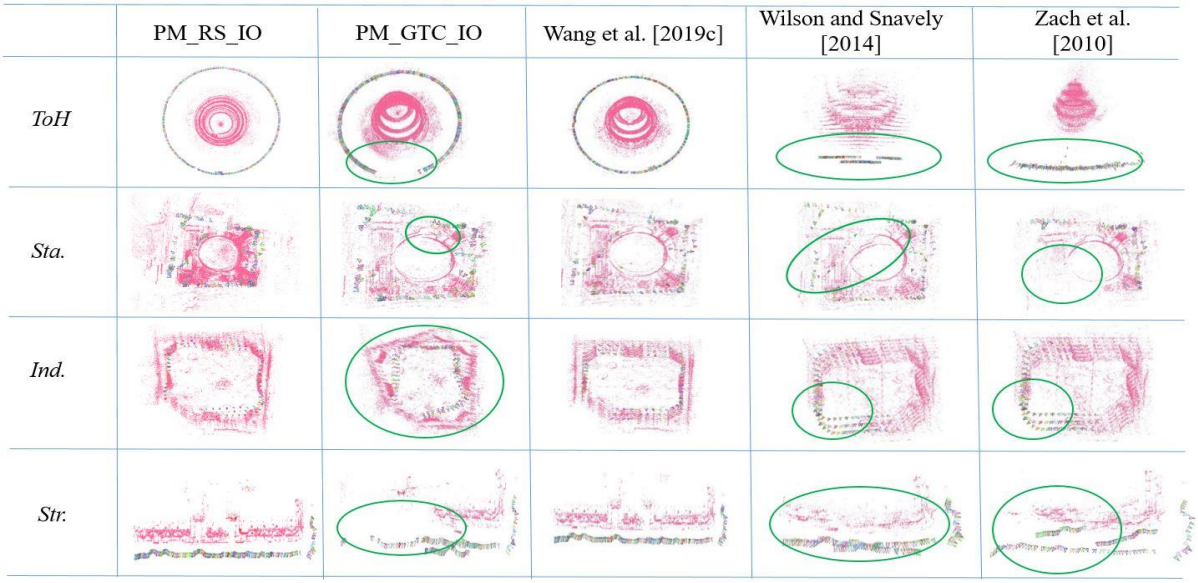
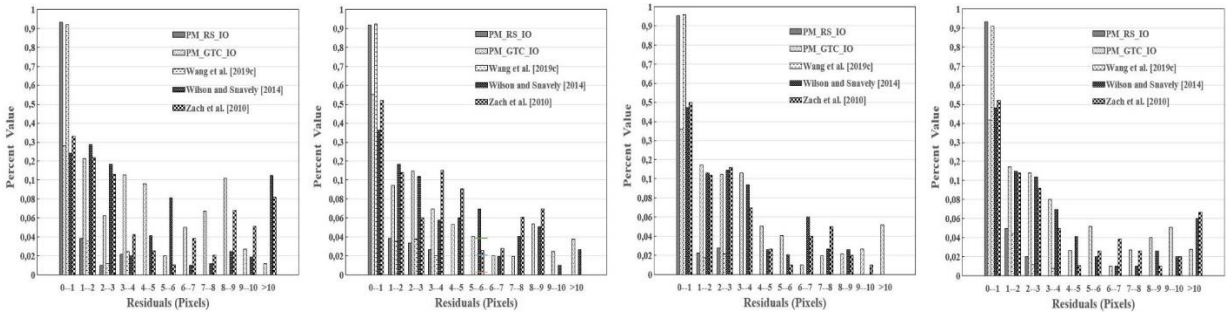


Figure 6.21: Visualization of reconstruction results on four public datasets from PM_RS_IO, PM_GTC_IO, Wang et al. [2019], Wilson and Snavely [2014] and Zach et al. [2010].

To further underline the conclusions from Figure 6.21, it is necessary to again show the reprojection error distribution for these four datasets (see Figure 6.22). Form this figure the same conclusions can be drawn as from Figure 6.20.



(a) *ToH* (b) *Sta* (c) *Ind.* (d) *Str.*

Figure 6.22: Reprojection error distribution on ToH, Sta, Ind. and Str. compared with other methods.

As it has already been mentioned, the strategy to select correct ROs by Wang et al. [2019c] might fail on some datasets due to the fixed number of selected ROs, being seven times the number of connected images after filtering. To explore this issue further, two additional datasets, namely, *Capitole* and *CAB*, are tested. Figure 6.23 shows the reconstruction results of these two datasets with ROs from different methods. Analogously to the results shown in Figure 6.21, Zach et al. [2010], Wilson and Snavely [2014] and PM_GTC_IO generate different artefacts as the ellipses indicate. Wang et al. [2019c] do improve the reconstruction results, as the artefacts are smaller, which is also illustrated by the reprojection error distribution shown in Figure 6.24. However, a comparison of the reconstruction results of PM_RS_IO and Wang et al. [2019c] reveals that the

suggested self-adapting method performs better than the one proposed by Wang et al. [2019c] (see Figures 6.23 and 6.24).

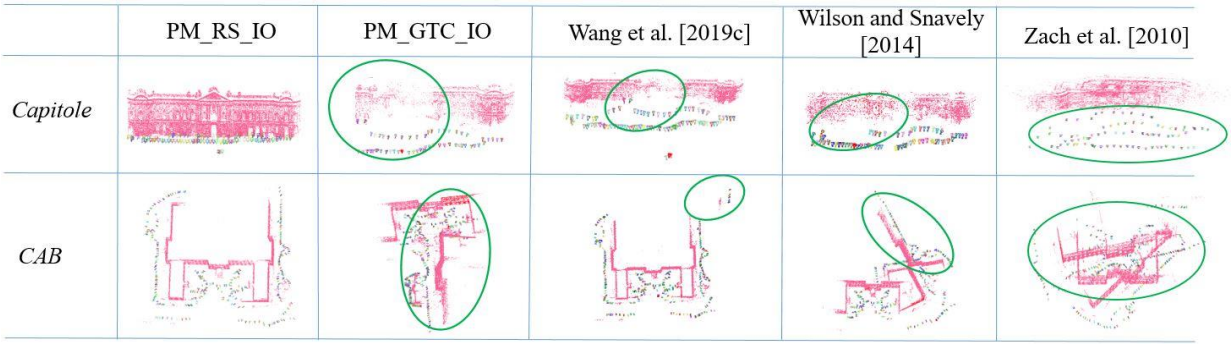
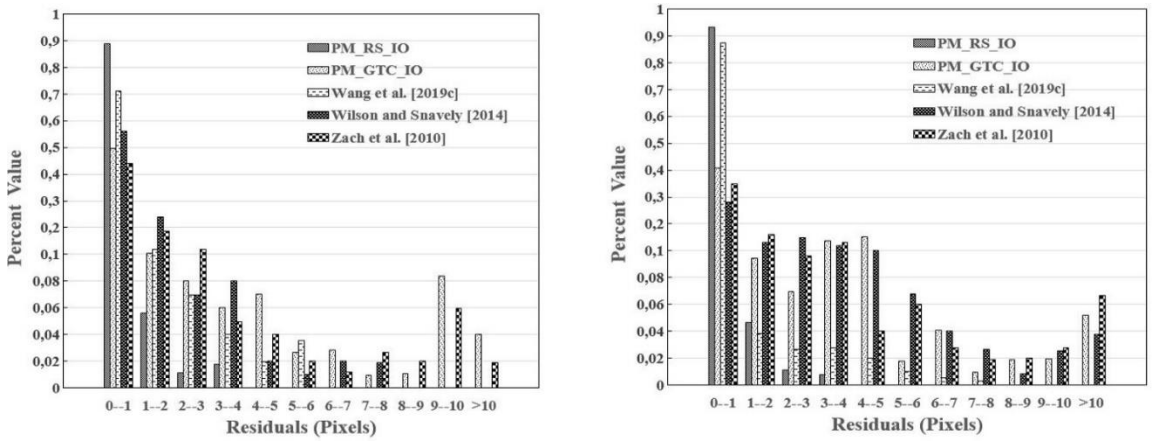


Figure 6.23: Visualization of reconstruction results on Capitole and CAB from PM_RS_IO, PM_GTC_IO, Wang et al. [2019c], Wilson and Snavely [2014] and Zach et al. [2010].

For a better analysis of these two reconstruction results, the reprojection error distributions of *Capitole* and *CAB* from different methods are shown in Figure 6.24. These results reveal that in 90% of the cases, the reprojection error of PM_RS_IO is less than one pixel, and the rest of them are spread over the ranges of 1-2, 2-3, 3-4 pixels. For the three other methods of Zach et al. [2010], Wilson and Snavely [2014] and PM_GTC_IO, in contrast, a much smaller percentage of reprojection errors are smaller than one pixel and residuals with larger values (e.g., 3-4, 4-5, 5-6 pixels) make up a relatively high percentage. Coming to Wang et al. [2019c], Figures 6.24 (a) and (b) imply that PM_RS_IO outperforms this approach, because more reprojection errors are assigned to ranges of smaller values.



(a) *Capitole* (b) *CAB*

Figure 6.24: Reprojection error distribution of Capitole and CAB compared with other methods.

In summary, based on these public datasets with highly repetitive structure, it turns out that PM_RS_IO provides the best reconstruction results compared to the other methods, i.e., Wang et al. [2019c], Zach et al. [2010], Wilson and Snavely [2014] and PM_GTC_IO.

Datasets with BPVD

To investigate the presented method on identifying correct BPVD ROs and the strategy of selecting robust correspondences from these correct BPVD image pairs, this section reports the results on four BPVD datasets; some sample images are shown in Figure 5.4. For these close-range

datasets, two different scenarios are considered, namely, outdoor and indoor scenes, *CR1* is the outdoor image set capturing images of a building's facades and *CR2* is the indoor image set taken along a narrow corridor. The outdoor scenes were captured using a UAV - first some images of a built-up area were taken from various flying heights, then, images of a planar grassland patch were captured. Since there is no repetitive structure in these datasets, the function for eliminating RS ROs is turned off.

Similar to the experiments mentioned before, the suggested global image orientation method is used for recovering the image poses and the coordinates of object points, the results are shown in Figure 6.25. To demonstrate the proposed method's performance on BPVD datasets, the reconstruction results of the proposed global image orientation method without any ROs blunder detection (denote as PM_IO) and with BPVD ROs processing (denote as PM_BPVD_IO) are compared. From Figure 6.25, it can be seen that in the PM_IO case all four BPVD datasets generated different visual artefacts shown by the solid ellipses. PM_BPVD_IO generates a better result, specifically, it does not generate so many object points somewhere in the air for *CR1*, no motion drift happens in *CR2*, and the generated object point cloud does not show a somewhat convex shape for the two UAV image sets. To improve the robustness of the suggested global image orientation method, many object points around the viewing direction (close to the principal points in image space) are eliminated before reconstruction as the dashed ellipses illustrate, this is due to the employed correspondence selection strategy described in Section 3.2.4. Note that the photogrammetric block does not contain any image if only the VSB criterion is used, because all image pairs of BPVD will be eliminated as RO outliers due to VSB (the corresponding reconstruction result is not shown here, because there is no reconstruction at all).

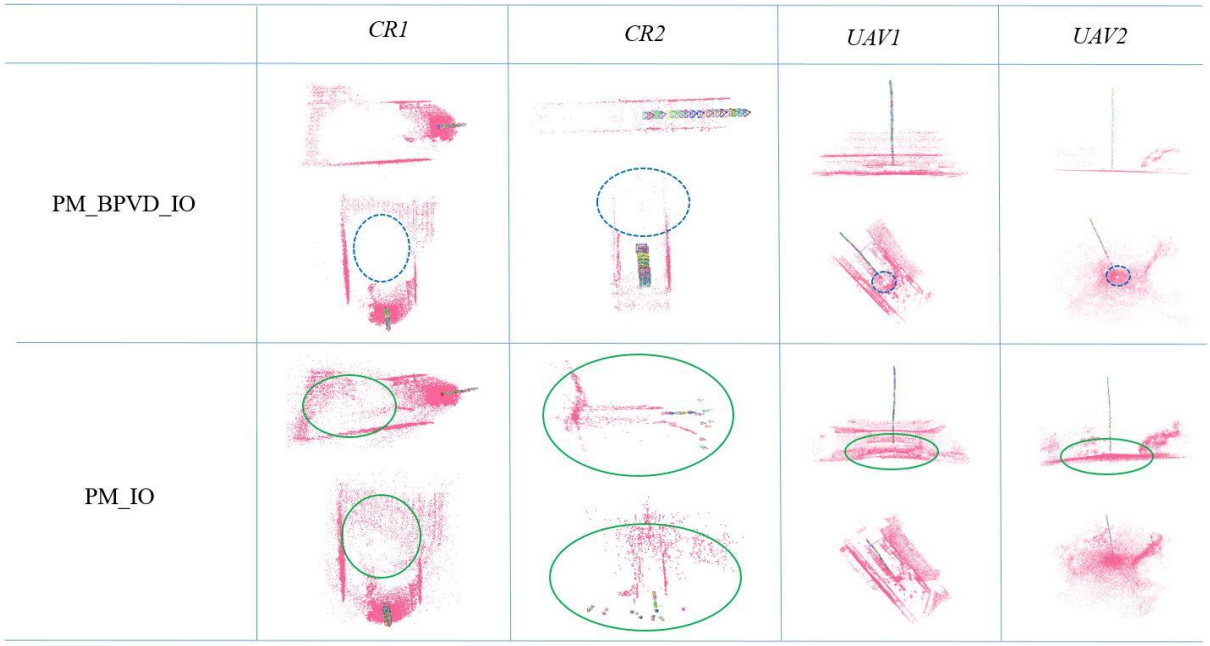
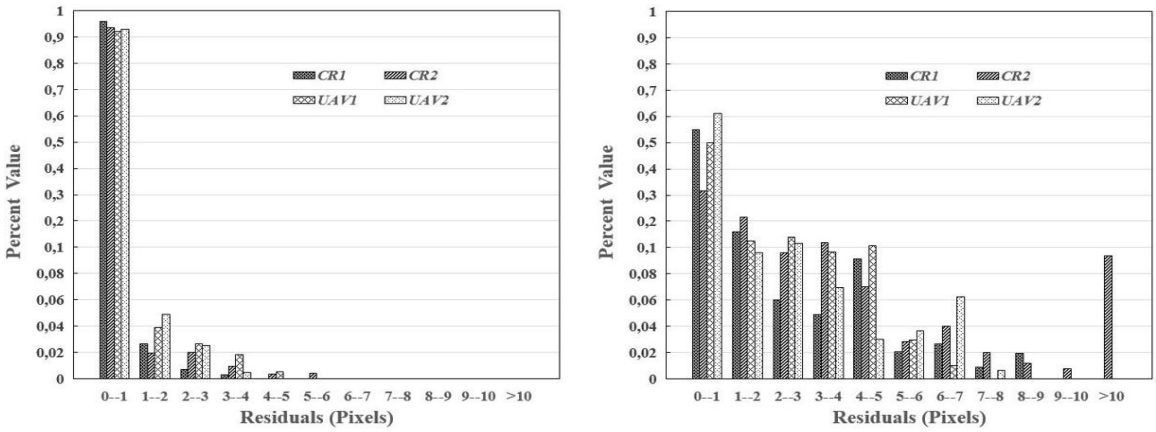


Figure 6.25: Visualization of reconstruction results on BPVD datasets. Dashed ellipses denote areas which are near the viewing direction, solid ellipses indicate visual artefacts.

Figure 6.26 provides numerical results for the self-generated BPVD datasets. Analyzing Figure 6.26 (a) and (b) it can be seen that the proposed method to increase the robustness of ROs has a

positive influence on the presented global image orientation method. In detail, for all four BPVD datasets more than 90% of the reprojection errors are smaller than 1 pixel and the maximum residuals are all below 5 pixels (except for *CR2* where it is below 6 pixels).



(a) PM_BPVD_IO (b) PM_IO

Figure 6.26: Reprojection error distribution of self-generated BPVD datasets.

Challenging dataset and Complex dataset

To further explore the potential of the proposed methods, two further datasets, namely, *Quad* [Crandall et al., 2011] and *Church* [Michelini and Mayer, 2020], are evaluated. *Quad* is denoted as challenging dataset because many global image orientation methods failed to deal with it due to RS (see the corresponding sample image in Figure 5.5) and image pairs with critical geometric configuration such as VSB. The complexity of *Church* results from the image configurations including wide baselines as well as terrestrial and UAV images with significantly different viewing direction. In addition, also blunder ROs due to both RS and critical configurations of VSB and BPVD exist in this dataset.

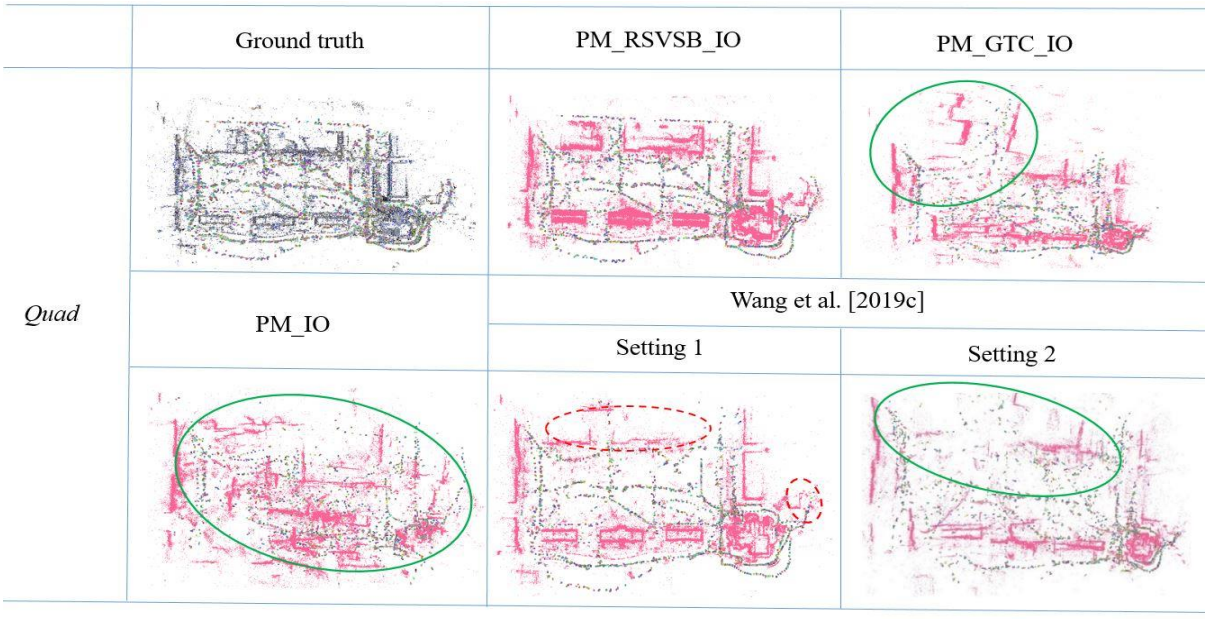


Figure 6.27: Visualization of reconstruction results of Quad for different methods. Red dashed ellipses denote areas where some images are missing. Green solid ellipses depict visual artefacts.

First, the performance on the *Quad* dataset is inspected using various methods, with Figure 6.27 visualizing the corresponding reconstruction results. Wang et al. [2019c] was implemented with two sets of free parameters provided by the authors (Setting 1 is adjusted to generate a better result for *Quad*, while Setting 2 is the default setting). Compared to the ground truth of the reconstruction [Crandall et al., 2011], the suggested PM_RSVSB_IO obtains the most reasonable result. Setting 1 of Wang et al. [2019c] loses some images (as the dashed ellipses show). Lots of visual artefacts occur in the results of Setting 2 of Wang et al. [2019c]. Similar results were obtained for the original ROs without any robustification and for ROs after the outlier elimination using the general rotation and translation triplet compatibility check. This particularly illustrates the difficulties of this dataset.

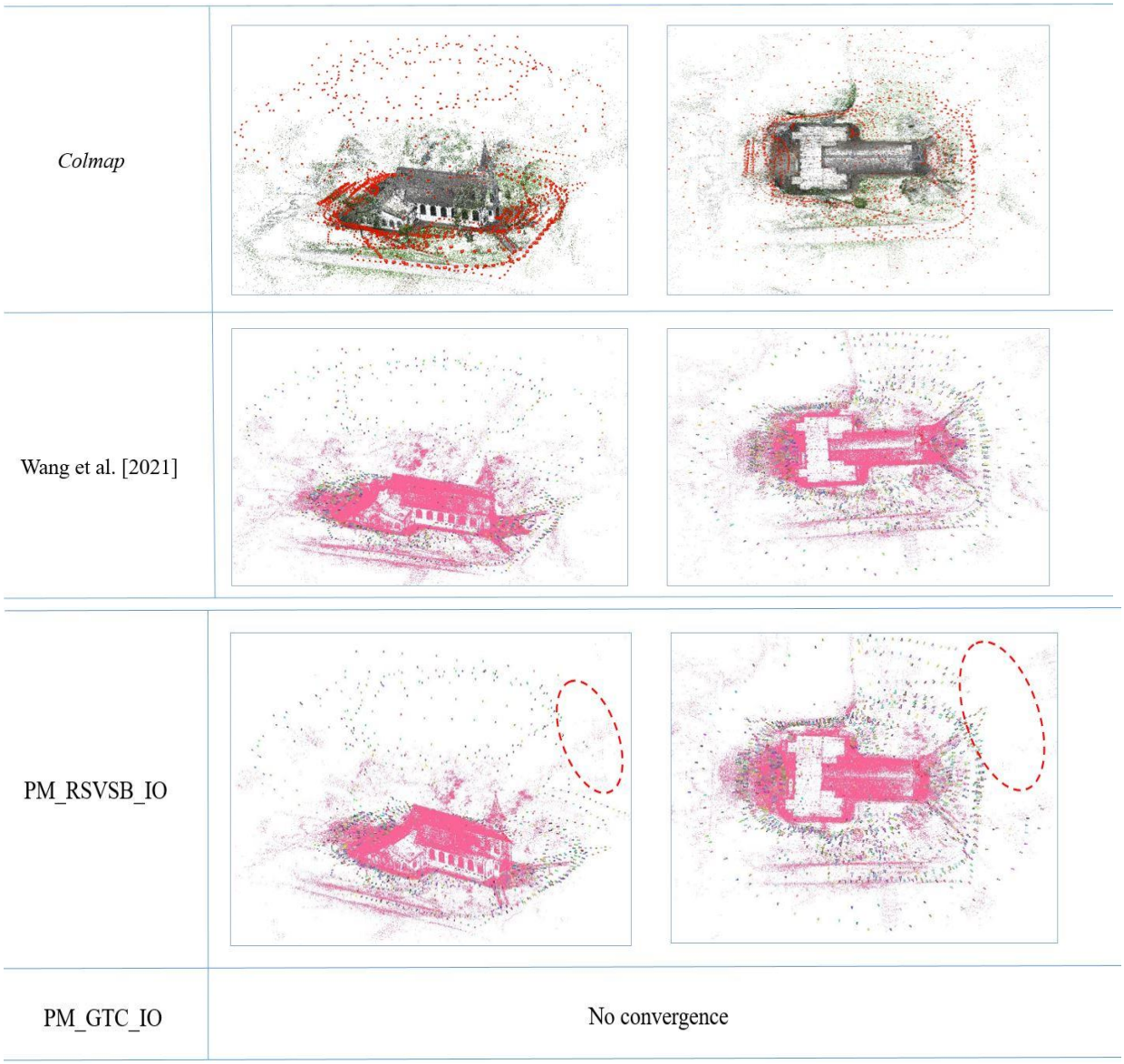


Figure 6.28: Visualization of reconstruction results of Church using various pipelines. For each pipeline, two pictures from different perspectives are shown. Red dashed ellipses denote areas where some images are missing. Green solid ellipses depict artefacts.

	N_r	Number of object points	σ_0 (in pixels)	Runtime of image orientation (in seconds)	Iterations of final bundle adjustment
Michelini and Mayer [2020]	/	290748	0.55	**396**	/
Colmap	**1444**	**550962**	0.97	16762	5
Wang et al. [2021]	1420	440962	**0.42**	1334	11
PM_RSVSB_IO	1405	542635	0.46	1663	13
PM_GTC_IO	-	-	-	-	No convergence

Table 6.17: Comparison of Church from different pipelines. σ_0 is the mean reprojection error. The results of Colmap[5] (Schönberger and Frahm, 2016), Wang et al. [2021], PM_RSVSB_IO and PM_GTC_IO were executed on the same machine as previous experiments used, namely, a quad-core processor (3.2 GHz Inter (R) Core (TM)i5-6500, 32G memory) and eight threads in total. The result of Michelini and Mayer [2020] is directly cited from their paper, which was generated on a machine of 2×Intel® Xeon® E5-2643 v3 (6 cores, 3.40 GHz). The best result in each column appears in bold. "/" means the corresponding item is not available. "-" means that the corresponding item is not provided due to the non-converged refinement of bundle adjustment.

Next, the reconstruction results of the complex dataset (*Church*) are reported. As no reference is available, a relatively coarse quantitative comparison together with a qualitative evaluation are given only and shown in Table 6.17 and Figure 6.28: Compared with *Colmap*, the number of reconstructed object points and orientated images is slightly lower for the proposed PM_RSVSB_IO. However, this method performs much better than *Colmap* with a lower mean reprojection error and an accelerated runtime (10.2 times faster). Inspecting the performance of Wang et al. [2021] in comparison to PM_RSVSB_IO, less time is used to generate a better (regarding the mean reprojection error) but sparser reconstruction. Figure 6.28 shows reconstruction results of these four pipelines, assuming the reconstruction of *Colmap* to be the reference as 1444 out of 1455 images are solved. Both the presented PM_RSVSB_IO and Wang et al. [2021] can reconstruct the church as a whole, yet, some images are missing in the results of PM_RSVSB_IO as the red dash ellipses indicate. As it has already been clarified, the general ROs elimination method using the triplet compatibility constraints cannot cope with RS and VSB ROs, this leads to that the bundle adjustment dose not converge when using the initial image orientation parameters estimated by the PM_GTC_IO method (therefore, the qualitative and quantitative results are not shown). In particular, the iteration number of bundle adjustment in PM_GTC_IO exceeds the maximum threshold (50 times) while the bundle adjustment of other two global methods converge after about a dozen iterations (11 iterations for Wang et al. [2021] and 13 iterations for PM_RSVSB_IO). Compared to Michelini and Mayer [2020], the most attractive virtue of the proposed PM_RSVSB_IO pipeline is that up to 250,000 object points are additionally reconstructed with an even smaller mean reprojection error. However, the method of Michelini and Mayer [2020] outperforms the other approaches with respect to time efficiency (4.2 times faster than PM_RSVSB_IO), which is mainly caused by two factors: first, the significantly lower number of object points reduces the runtime of the bundle adjustment; second, parallelization is utilized on a more powerful machine (2 × Intel® Xeon® E5-2643 v3 (6 cores, 3.40 GHz)) by them.

[5] The package is available at https://colmap.github.io, Version 3.2, the corresponding default settings are applied.

6.3 Synthesis

The presented methods are thoroughly assessed by the comprehensively conducted experiments that are described in the previous subsections. In this section, the most remarkable findings of these experiments are synthesized.

6.3.1 Preprocessing steps

In this thesis, preprocessing concentrates on solving two main issues regarding time efficient image matching and robustification of the input ROs.

Time efficient image matching

To demonstrate the performance of the presented fast image matching approach, two types of datasets consisting of ordered and unordered internet datasets are tested, which are considered to be representative. Based on the exhaustive pairwise image matching, a relevant efficacy of over 90 percent precision and over 50 percent recall after epipolar geometric validation can be achieved around 3 to 44 times faster using the proposed method. Compared with the method VocMatch, approximately the same runtime is achieved, while reaching a higher precision and recall.

Robustification of the input ROs

The effectiveness of the ROs robustification method is illustrated, while validating the general blunder ROs elimination method checking the triplet compatibility on *Lejonet* (ordered internet dataset) and *Piazza del Popolo* (unordered internet dataset). It is shown that the most of remaining ROs are correct and that most of the eliminated ROs are indeed incorrect. In addition, three benchmarks with ground truths ROs are employed to evaluate the proposed method's capability of dealing with RS and VSB ROs: more than 90 percent of the detected ROs are correct and more than 90 percent of the truly correct ROs are detected as well. Furthermore, the corresponding experimental results reveal that using the ROs robustification method improves the convergence behavior of the global rotation estimation method [Chatterjee and Govindu, 2013] with respect to the convergence speed and to the accuracy of the final solution. Finally, also less inaccurate object points are produced after increasing the robustness of the ROs.

6.3.2 Global image orientation

The quality of the estimations resulting from global image orientation are highly dependent on the inputs generated by preprocessing steps. In order to evaluate the accomplishment of the research objectives described in Chapter 1, various datasets are orientated and evaluated. Employing ordered and unordered datasets using different methods, the incremental method shows superior results with respect to accuracy, while the proposed global image orientation method reaches

almost the same level of accuracy after bundle adjustment, while typically being 2-20 times faster. The translation accuracy before and after bundle adjustment are studied, implying that the proposed method does not reach the same level of accuracy that is achieved after bundle adjustment. However, this method yields a good initialization for bundle adjustment. In addition, extensive experiments on abundant problematic datasets are conducted. Combined evaluations consisting of quantitative and qualitative results show that the proposed global image orientation method including the proposed ROs robustification method can provide a robust and reliable reconstruction result for most scenarios.

7 Conclusion and Outlook

This chapter closes the presented thesis and draws conclusions from the presented methods and the corresponding experiments with respect to the pursued objective. Moreover, several promising directions for potential future works are outlined.

The main research goal of this work is to develop *a novel method for determining accurate exterior orientation parameters in a fast and robust manner*. This goal was achieved by two continuous lines of research: *preprocessing* consisting of time efficient image matching and ROs robustification, and *global image orientation* consisting of global rotation and translation estimation. 'Fast' was approached by accelerating the procedure of image matching and image orientation which are typically the two most time-consuming steps. 'Robust' was addressed by the detection and elimination of RO outliers and the used robust bundle adjustment.

Preprocessing steps can be treated as tools for yielding accurate inputs for the subsequent image orientation. Two main issues were addressed at this stage: image matching and handling outliers in ROs:

1. First, a method based on random k-d forest was developed for fast seeking mutual overlapping image pairs. In particular, features from all images were employed to build several k-d trees that were supposed to be as independent to each other as possible. Similarity degree values among potential pairs were then estimated for identifying mutual overlapping pairs. Without any prior information, time efficiency is improved by only conducting image matching and epipolar geometry validation on these determined overlapping image pairs. According to the tests on dozens of ordered and unordered internet datasets, the efficacy is discussed by considering the results of exhaustive pairwise image matching as reference. After epipolar geometry computation, the precisions and recalls are nearly all over 90 and 50 percent, respectively, on both unordered and ordered internet datasets. These values have been shown to be good enough for executing image orientation. Furthermore, the time efficiency is improved by a factor ranging from 2 to 41 times depending on the size of the specific dataset.
2. Second, in order to improve the robustness of image orientation, approaches for the robustification of ROs were presented. Other than verifying the geometric relationships between image pairs, a general method was proposed by checking the compatibility among triplets. For this purpose, two different compatibility measures were estimated using relative rotations and relative translations. It was shown that the remaining ROs were indeed less contaminated by RO outliers regarding to the relative rotation and translation error. In addition, a method was designed to cope with wrong ROs resulting from repetitive structure (RS) and very short baselines (VSB). Criteria for these two cases were introduced based on the number

of conjugate points in overlapping pairs and on the magnitude of the intersection angle during triangulation. As the latter criterion is sensitive to pairs with a long enough baseline approximately parallel to the viewing direction, a new criterion was presented accordingly. The corresponding performance was quantitatively and qualitatively demonstrated on various datasets, and the positive influences on the global rotation estimation method [Chatterjee and Govindu, 2013] and on the reconstructed object points were further shown.

In this thesis, a two-step *global image orientation* method consisting of global rotation estimation and global translation estimation is advocated. As ample research has been published on solving global rotations from pairwise relative rotations, this thesis concentrates on determining global translations. To make this work more self-contained, the global rotation estimation method by Chatterjee and Govindu [2013] was nevertheless presented. It was shown that the presented ROs robustification method can improve the approach of Chatterjee and Govindu [2013] with respect to both the accuracy of the solution and the convergence rate. In addition, a new method that can simultaneously estimate translations for all available images was proposed. In this context, globally consistent scale factors for every remaining relative translation were computed using the depth of tie points from individual local spatial intersection. The corresponding global translations were then determined by averaging these scaled relative translations. Finally, these estimated initial values were refined by the introduced robust bundle adjustment.

The performance of global image orientation was thoroughly demonstrated on various datasets: the reconstruction results of ordered and unordered datasets were compared with one conventional incremental SfM method and several state-of-the-art global SfM methods. Compared to the incremental method, approximately the same accuracy was achieved by the presented global method, while the runtime was reduced by a factor around 10 – 20. All the investigated global image orientation methods achieved nearly the same accuracy and time efficiency, whereas, the proposed global method was superior when dealing with unordered datasets. To further prove that the presented ROs robustification method is beneficial for the presented global image orientation method, various problematic datasets were tested. Datasets containing challenging cases of repetitive structure (RS), very short baselines (VSB) and baseline parallel to the viewing direction (BPVD) were reconstructed using the global image orientation method and the uncontaminated ROs generated by various methods. The results show that the proposed method integrated with the corresponding proposed RO robustification method clearly produces the best reconstruction result. To further explore the presented approaches' potentials, one challenging dataset (6514 images) and one complex dataset (1455 images) were tested, in addition. The proposed method was able to reconstruct both datasets with very good results.

There are several future directions which can be followed up to further improve the present approaches.

First, threshold investigation. In this thesis, the implemented thresholds as introduced in Table 5.1 were empirically selected without fully detailed trials. Although the empirical selection was demonstrated to be efficient on all experimental datasets, future efforts could be paid on this issue to try to learn better settings via testing more different threshold settings and datasets. For example,

from the presented qualitative and quantitative results, the reconstruction as a whole could be generally recovered while few images were still missing which may lead to a broken model.

Second, gauge definition. Since all the relevant problems stated in this thesis are still in the frame of a free network, the gauge definition issue exists in both global rotation estimation and global translation estimation, i.e., the first image was normally selected to have a 3×3 identity rotation matrix and a zero vector for translation. However, this may not be an optimal choice, for example, the risk of a degenerated solutions is increased if the first image is weekly connected to the photogrammetry block. As Wilson et al. [2016] suggested, one possible way which deserves further research is that an image with more connected images should have a higher priority to define the gauge.

Finally, bundle adjustment. While bundle adjustment is not part of the contribution of this work, bundle adjustment is always the most time-consuming procedure in estimating orientation parameters. Thus, several extensions can be carried out to further accelerate this procedure. First, tie point selection. Typically, bundle adjustment refines all the generated tie points which are usually noisy and redundant. Thus, it might be interesting to only select a minimal set of best tie points. Such an approach could extremely reduce the computational burden of bundle adjustment, because the number of tie points is normally much larger than the number of images. Second, applying GPU techniques. Applying graphic hardware GPU on bundle adjustment is not a very recent idea, since it was first implemented by Wu et al. [2011]. In this work, the preconditioned conjugate gradients were integrated with the GPU to quickly solve a very large linear equation system, but, this approach is limited due to the missing information of the solution’s uncertainty. Third, distributed strategies. In order to efficiently conduct the optimization task for a very large dataset, the large bundle adjustment optimization can be distributed into various subtasks and be carried out on communicating branch machines (such as, Zhang et al. [2017] and Mayer [2019]), whereas the final result is synthesized by a master machine.

In summary, the expected objectives of this thesis have been successfully met. Starting from the obtained features, the proposed method can be treated as a black box for extracting the corresponding image orientation parameters. According to the comprehensively reported experiments, this black box is applicable to fast and robustly orient image datasets and yet can be further updated by using a more reasonable threshold setting and gauge definition, or via a more advanced bundle adjustment pipeline.

Appendix

A. Proposition for very short baselines

The elements of the relative rotation matrix *R* can be accurately estimated from the essential matrix, no matter how short the baseline length is.

Proof. Inspired by the calculation of the essential matrix [Hartley and Zisserman, 2003], the correspondences are employed to obtain a solution of a 3×3 matrix $\boldsymbol{L}$.

$$\boldsymbol{x}_j{}^T \boldsymbol{L} \boldsymbol{x}_i = 0 \Leftrightarrow (\boldsymbol{x}_i{}^T \otimes \boldsymbol{x}_j{}^T)\, vec(\boldsymbol{L}) = 0 \tag{A.1}$$

where ⊗ denotes the Kronecker product, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the coordinates of all the correspondences from image *i* and *j*, *vec*(.) is the vectorization of a matrix. Besides this, it is also feasible to get a formula $\boldsymbol{X_P} = \boldsymbol{X_{Ci}} + \lambda_i \boldsymbol{R_i x_i}$ and $\boldsymbol{X_P} = \boldsymbol{X_{Cj}} + \lambda_j \boldsymbol{R_j x_j}$, where $\boldsymbol{X_P}$ denotes coordinate vector of object point *P* and $\boldsymbol{X_{Ci}}$, $\boldsymbol{X_{Cj}}$ are projection centers of image *i* and *j* in the object space. λ_i and λ_j are the corresponding scale factors, $\boldsymbol{R_i}$ and $\boldsymbol{R_j}$ are the corresponding rotations from image to object space.

$$\boldsymbol{X_P} = \boldsymbol{X_{Ci}} + \lambda_i \boldsymbol{R_i x_i} = \boldsymbol{X_{Cj}} + \lambda_j \boldsymbol{R_j x_j} \tag{A.2}$$

This can be rewritten as

$$\boldsymbol{x_j} = \lambda_{ij} (\boldsymbol{R_{ij}}\, \boldsymbol{x_i} + v_i \boldsymbol{t_{ij}}) \tag{A.3}$$

$$\lambda_i \boldsymbol{R_i x_i} = \boldsymbol{X_{Cj}} - \boldsymbol{X_{Ci}} + \lambda_j \boldsymbol{R_j x_j} \tag{A.4}$$

where $\boldsymbol{t_{ij}} = \boldsymbol{R_j}^{-1}(\boldsymbol{X_{Ci}} - \boldsymbol{X_{Cj}})$ is the baseline vector and $\boldsymbol{R_{ij}} = \boldsymbol{R_j}^{-1} \boldsymbol{R_i}$ is the relative rotation, $\lambda_{ij} = \lambda_i / \lambda_j$, $v_i = 1/\lambda_i$.To take the relative rotation and translation into consideration, (A.1) is rewritten by using the mixed-product property of ⊗ and equation (A.3),

$$(\boldsymbol{x}_i{}^T \otimes (\lambda_{ij} (\boldsymbol{R_{ij}}\, \boldsymbol{x}_i + \boldsymbol{v}_i \boldsymbol{t_{ij}})^T)\, vec(\boldsymbol{L}) = 0$$

$$\Leftrightarrow (\boldsymbol{x}_i{}^T \otimes [\boldsymbol{x}_i{}^T \;\; \boldsymbol{v}_i])\, (I_3 \otimes [\boldsymbol{R_{ij}} \quad \boldsymbol{t_{ij}}]^T) vec(\boldsymbol{L}) = 0 \tag{A.5}$$

in which λ_{ij} is eliminated as λ_{ij} is always larger than 0, $\boldsymbol{v}_i$ contains all the v_i values from all the correspondences.

$$\begin{Bmatrix} \boldsymbol{x_1}^T \otimes [\boldsymbol{x_1}^T \ \ v_1\,] \\ \boldsymbol{x_2}^T \otimes [\boldsymbol{x_2}^T \ \ v_2\,] \\ \vdots \\ \boldsymbol{x_m}^T \otimes [\boldsymbol{x_m}^T \ \ v_2\,] \end{Bmatrix} (\boldsymbol{I}_3 \otimes [\boldsymbol{R_{ij}} \quad \boldsymbol{t_{ij}}]^T)\, vec(\boldsymbol{L}) = 0 \tag{A.6}$$

Thus, equation (A.6) is obtained, where m is the number of correspondences. Then, $\boldsymbol{U}$ represents the parameter matrix and $\boldsymbol{z}$ representing the unknowns, namely $\boldsymbol{R_{ij}}$, $\boldsymbol{t_{ij}}$ and L

$$\boldsymbol{U} = \begin{Bmatrix} \boldsymbol{x_1}^T \otimes [\boldsymbol{x_1}^T \ \ v_1\,] \\ \boldsymbol{x_2}^T \otimes [\boldsymbol{x_2}^T \ \ v_2\,] \\ \vdots \\ \boldsymbol{x_m}^T \otimes [\boldsymbol{x_m}^T \ \ v_2\,] \end{Bmatrix}, \ \boldsymbol{z} = (\boldsymbol{I}_3 \otimes [\boldsymbol{R_{ij}} \quad \boldsymbol{t_{ij}}]^T)\, vec(\boldsymbol{L}) \tag{A.7}$$

$$\boldsymbol{U}\boldsymbol{z} = 0 \tag{A.8}$$

Via analyzing the $\boldsymbol{U}$ matrix and unfolding it,

$$\boldsymbol{U} = \begin{Bmatrix} x_1^2 & x_1y_1 & x_1 & x_1v_1 & y_1x_1 & y_1^2 & y_1 & y_1v_1 & x_1\, y_1 & 1 & v_1 \\ x_2^2 & x_2y_2 & x_2 & x_2v_2 & y_2x_2 & y_2^2 & y_2 & y_2v_2 & x_2\, y_2 & 1 & v_2 \\ & & & & & \vdots & & & & & \\ x_m^2 & x_my_m & x_m & x_mv_m & y_mx_m & y_m^2 & y_m & y_mv_m & x_m\, y_m & 1 & v_m \end{Bmatrix} \tag{A.9}$$

It can be found that columns 2, 3 and 7 are equal to columns 5,9 and 10, so, $rank(\boldsymbol{U}) \leq 9$. Therefore, when $m \geq 9$, the homogeneous equation (A.8) has three linearly independent basic solutions

$$\begin{aligned} \boldsymbol{\varepsilon_1} &= (0\ 1\ 0\ 0 -1\ 0\ 0\ 0\ \ 0\ \ 0\ 0\ 0\,) \\ \boldsymbol{\varepsilon_2} &= (0\ 0\ 1\ 0\ \ 0\ \ 0\ 0\ 0\ -1\ \ 0\ 0\ 0\,) \\ \boldsymbol{\varepsilon_3} &= (0\ 0\ 0\ 0\ \ 0\ \ 0\,1\ \ 0\ \ 0 -1\,0\ 0\,) \end{aligned} \tag{A.10}$$

Thus, the general solution space of z is

$$\boldsymbol{z} = (\boldsymbol{I}_3 \otimes [\boldsymbol{R_{ij}} \quad \boldsymbol{t_{ij}}]^T)\, vec(\boldsymbol{L}) = (k_1\boldsymbol{\varepsilon_1} + k_2\boldsymbol{\varepsilon_2} + k_3\boldsymbol{\varepsilon_3}) \tag{A.11}$$

in which k_1, k_2, k_3 are all real numbers. According to (A.7),

$$(\boldsymbol{I}_3 \otimes [\boldsymbol{R_{ij}} \quad \boldsymbol{t_{ij}}]^T)\, vec(\boldsymbol{L}) = vec([\boldsymbol{R_{ij}} \quad \boldsymbol{t_{ij}}]^T \boldsymbol{L}) \tag{A.12}$$

Substituting (A.12) in equations (A.11) yields

$$[\boldsymbol{R_{ij}} \quad \boldsymbol{t_{ij}}]^T \boldsymbol{L} = \begin{Bmatrix} \boldsymbol{R_{ij}}^T \boldsymbol{L} \\ \boldsymbol{t_{ij}}^T \boldsymbol{L} \end{Bmatrix} = \begin{Bmatrix} 0 & -k_1 & -k_2 \\ k_1 & 0 & -k_3 \\ k_2 & k_3 & 0 \\ 0 & 0 & 0 \end{Bmatrix} \tag{A.13}$$

Then, the relationships between $\boldsymbol{R_{ij}}$ and $\boldsymbol{L}$, $\boldsymbol{t_{ij}}$ and $\boldsymbol{L}$ are formulated as

$$\boldsymbol{L} = \boldsymbol{R_{ij}} \begin{Bmatrix} 0 & -k_1 & -k_2 \\ k_1 & 0 & -k_3 \\ k_2 & k_3 & 0 \\ 0 & 0 & 0 \end{Bmatrix} = \boldsymbol{R_{ij}}[\boldsymbol{k}]_{\odot}, \quad [\boldsymbol{k}]_{\odot} = \begin{Bmatrix} 0 & -k_1 & -k_2 \\ k_1 & 0 & -k_3 \\ k_2 & k_3 & 0 \\ 0 & 0 & 0 \end{Bmatrix} \tag{A.14}$$

$$\boldsymbol{t_{ij}}^T L = 0 \tag{A.15}$$

and $\boldsymbol{k} = (k_1, -k_2, k_3)$. (A.14) means that the solution L of (A.1) is the essential matrix [Hartley et al., 2013]. So, $\boldsymbol{R_{ij}}$ can be decomposed from $\boldsymbol{L}$. From (A.14) and (A.15), when $\boldsymbol{t_{ij}} \neq \boldsymbol{0}$, $\boldsymbol{t_{ij}} = \pm \boldsymbol{R_{ij}}\boldsymbol{k}$ and $\boldsymbol{R_{ij}}$ can be correctly estimated using SVD decomposition [Hartley and Zisserman, 2003]. For $\boldsymbol{t_{ij}} = 0$, it is clear that L has no relationship with $\boldsymbol{t_{ij}}$, and it is still related to $\boldsymbol{R_{ij}}$ and $\boldsymbol{k}$. $\boldsymbol{k}$ can never be a zero vector which means $\boldsymbol{t_{ij}}$ is not related to $\boldsymbol{k}$, because the solution z will be zero if $\boldsymbol{k}$ is a zero vector and this requires that the homogenous equation has a full rank which means $Rank(U) = 12$, and this can never happen. Therefore, $\boldsymbol{R_{ij}}$ can still be correctly estimated from L when $\boldsymbol{t_{ij}} = 0$, and the corresponding solution for $\boldsymbol{t_{ij}}$ is not the correct relative translation, but the $\boldsymbol{k}$ vector.

B. Calculation of the discrepancy between relative orientation and ground truth exterior orientation parameters

B.1 Discrepancy with respect to relative rotations

Given one relative rotation $\widehat{\boldsymbol{R}}_{\boldsymbol{ij}}$ for images i and j, estimated by the decomposition of an essential matrix using the five-point method, and the corresponding ground truth global rotation $\boldsymbol{R_i}$ and $\boldsymbol{R_j}$, then, the discrepancy of relative rotation from the ground truth can be computed by

$$\beta = \left(\text{trace}(\widehat{\boldsymbol{R}}_{\boldsymbol{ij}}^{\boldsymbol{-1}} \boldsymbol{R_i} \boldsymbol{R_j^{-1}}) - 1\right) / \boldsymbol{2} \tag{B.1}$$

$$\theta_r = \arccos(\beta) \cdot 180 / \boldsymbol{\pi} \tag{B.2}$$

where β is the average value of the main diagonal elements of $\widehat{\boldsymbol{R}}_{\boldsymbol{ij}}^{\boldsymbol{-1}} \boldsymbol{R_i} \boldsymbol{R_j^{-1}}$, and θ_r is the discrepancy value with respect to the relative rotations.

B.2 Discrepancy with respect to relative translations

As the length of a relative translation is normalized to 1 when decomposing the essential matrix, whereas, ground truth translation parameters are typically attributed with global scales, in the thesis, the corresponding discrepancy is computed as the intersection angle between the relative translation estimated from essential matrix and relative translation computed by the corresponding ground truth translation parameters. In principal, the more accurate the estimated relative translation is, the smaller should be this intersection angle. Given relative translation $\boldsymbol{t_{ij}}$ of the two

images *i* and *j*, and the corresponding ground truth global rotation $\boldsymbol{R_i}$ and global translations $\boldsymbol{C_i}$ and $\boldsymbol{C_j}$,

$$\check{\boldsymbol{t}}_{ij} = \boldsymbol{R_i}(\boldsymbol{C_j} - \boldsymbol{C_i}) \tag{B.3}$$

Equation (B.3) computes the ground truth relative translations of images *i* and *j*; to obtain the corresponding discrepancy, the intersection angle is computed by equation (B.4)

$$\theta_t = \arccos(\boldsymbol{t_{ij}} \cdot \check{\boldsymbol{t}}_{ij} / (\|\boldsymbol{t_{ij}}\| \|\check{\boldsymbol{t}}_{ij}\|)) \cdot 180/\boldsymbol{\pi} \tag{B.4}$$

and θ_t is the discrepancy value with respect to the relative translations.

C. Calculation of the mean translation errors

The image projection centres from ground truth are regarded as control points and the projection centres calculated by our method are transferred into the coordinate system of these control points using the following 3D similarity transformation [Horn et al., 1988]:

$$\boldsymbol{d}_i = \lambda(\boldsymbol{R}\boldsymbol{C}_i + \boldsymbol{T}),\ i=1,2,3\ldots n \tag{C.1}$$

where $\boldsymbol{d}_i$ are the coordinates of the control points and $\boldsymbol{C_i}$ are the corresponding projection centre coordinates, *n* is the number of the control points. λ is the scale factor, $\boldsymbol{R}$ is a 3×3 rotation matrix and $\boldsymbol{T}$ is a three-dimensional translation vector. There are seven parameters to be estimated, namely one scale factor, three rotation angles for $\boldsymbol{R}$ and three translation parameters for $\boldsymbol{T}$. One corresponding pair consisting of a control point and the corresponding projection centre yields three equations, so at least three point pairs are necessary to determine the seven unknowns.

RANSAC is used to solve this problem and the solution that results in the smallest mean translation error is kept. In this thesis, 4096 RANSAC iterations are conducted. Three pairs of control points and the corresponding image projection centres are randomly selected. The centroids $\boldsymbol{d_c}$ and $\boldsymbol{C_c}$ are calculated by equation (C.2),

$$\boldsymbol{d_c} = (\boldsymbol{d_1} + \boldsymbol{d_2} + \boldsymbol{d_3})/3,\ \ \boldsymbol{C_c} = (\boldsymbol{C_1} + \boldsymbol{C_2} + \boldsymbol{C_3})/3 \tag{C.2}$$

where $\boldsymbol{d}_1$, $\boldsymbol{d}_2$ and $\boldsymbol{d}_3$ are control points which are randomly chosen, $\boldsymbol{C_1}$, $\boldsymbol{C_2}$ and $\boldsymbol{C_3}$ are the corresponding projection centres. Using the centroids determines reduced coordinates $\boldsymbol{d_{ci}}$ and $\boldsymbol{C_{ci}}$ for *i*= {1,2,3}:

$$\boldsymbol{d_{ci}} = \boldsymbol{d_i} - \boldsymbol{d_c}\ ,\ \boldsymbol{C_{ci}} = \boldsymbol{C_i} - \boldsymbol{C_c} \tag{C.3}$$

Then, the scale λ is estimated using equation (C.4):

$$\lambda = (\|\boldsymbol{d_{c1}}\| + \|\boldsymbol{d_{c2}}\| + \|\boldsymbol{d_{c3}}\|)/(\|\boldsymbol{C_{c1}}\| + \|\boldsymbol{C_{c2}}\| + \|\boldsymbol{C_{c3}}\|) \tag{C.4}$$

where ||.|| is the L_2 norm. Using the resultant value of λ, $\boldsymbol{C}_{ci}$, i= {1,2,3} are transferred, into a coordinate system which has the same scale as the one of the control points, see equation (C.5):

$$\boldsymbol{C}_{sci} = \lambda \boldsymbol{C}_{ci}. \tag{C.5}$$

To estimate the rotation matrix $\boldsymbol{R}$ and the translation $\boldsymbol{T}$, equation (C.6) is optimized using least squares adjustment,

$$\begin{aligned}
&\underset{\boldsymbol{R},\boldsymbol{T}}{\operatorname{argmin}} \sum_{i=1}^{3} \left\| \boldsymbol{d}_i - \boldsymbol{R}\boldsymbol{C}_{sci} - \lambda \boldsymbol{T} \right\|^2 \\
&= \sum_{i=1}^{3} \left\| (\boldsymbol{d}_{ci} + \boldsymbol{d}_c) - \boldsymbol{R}(\boldsymbol{C}_{sci} + \lambda \boldsymbol{C}_c) - \lambda \boldsymbol{T} \right\|^2 \\
&= \sum_{i=1}^{3} \left\| \boldsymbol{d}_{ci} - \boldsymbol{R}\boldsymbol{C}_{sci} + \boldsymbol{d}_c - \lambda \boldsymbol{R}\boldsymbol{C}_c - \lambda \boldsymbol{T} \right\|^2 \\
&= \sum_{i=1}^{3} \left\| \boldsymbol{d}_{ci} - \boldsymbol{R}\boldsymbol{C}_{sci} \right\|^2 = \sum_{i=1}^{3} \left(\boldsymbol{d}_{ci}^T \boldsymbol{d}_{ci} + \boldsymbol{C}_{sci}^T \boldsymbol{C}_{sci} - 2\boldsymbol{d}_{ci}^T \boldsymbol{R}\boldsymbol{C}_{sci} \right)
\end{aligned} \tag{C.6}$$

because the term $\boldsymbol{d}_c - \lambda \boldsymbol{R}\boldsymbol{C}_c - \lambda \boldsymbol{T}$ is constant, the centroid of $\boldsymbol{C}_c$ can also be transferred to $\boldsymbol{d}_c$ by (C.1) and $\boldsymbol{R}^T\boldsymbol{R}$ =**1**. To minimize (C.6), the term $2\boldsymbol{d}_{ci}^T \boldsymbol{R}\boldsymbol{C}_{sci}$ must be maximized, which is equal to maximizing trace($\boldsymbol{R}\boldsymbol{H}$), with $\boldsymbol{H} = \sum_{i=1}^{3} \boldsymbol{C}_{sci} \boldsymbol{d}_{ci}^T$. $\boldsymbol{H}$ is decomposed by SVD (singular value decomposition) into $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$. The diagonal entries of $\boldsymbol{\Lambda}$ are the singular values of $\boldsymbol{H}$; $\boldsymbol{U}$ and $\boldsymbol{V}$ contain the left-singular vectors and the right-singular vectors of **H**. Finally, $\boldsymbol{R}$ is equal to $\mathbf{U}\boldsymbol{V}^T$, $\boldsymbol{T}$= $(\boldsymbol{d}_c - \lambda \boldsymbol{R}\boldsymbol{C}_c)/\lambda$. Given the values of λ, $\boldsymbol{R}$ and $\boldsymbol{T}$, the error of each image projection centre with respect to the "ground truth" can be determined as,

$$e_i = \left\| \boldsymbol{d}_i - \lambda(\boldsymbol{R}\boldsymbol{C}_i + \boldsymbol{T}) \right\|_2 \tag{C.7}$$

where $\boldsymbol{d}_i - \lambda(\boldsymbol{R}\boldsymbol{C}_i + \boldsymbol{T})$ is the translation error vector, it is normalized to obtain the position centre error e_i by using the L_2 norm. The mean position centre error $\bar{e}$ is calculated by averaging these normalized position errors.

References

Abdel-Aziz Y I, Karaa H M (1971) Direct linear transformation from comparator coordinates into object-space Coordinates. *ASP Symp. Close-Range Photogrammetry*, University of Illinois, Urbana, I11: 1-18.

Albertz J (2001) Albrecht Meydenbauer - Pioneer of Photogrammetric documentation of the cultural heritage. In: *Proceeding of 18th Int. Symposium CIPA.*

Albertz J (2009) 100 Jahre Deutsche Gesellshaft für Photogrammetrie. Fernerkundung und Geoinformation. *Photogrammetrie – Fernerkundung – Geoinformation*, 6: 487-560.

Agarwal S, Snavely N, Simon I, Seitz S and Szeliski R (2009) Building Rome in a day. In: *Proceedings of the IEEE International Conf. on Computer Vision (ICCV):* 72–79.

Agarwal S, Mierle K, et al. (2007): Ceres Solver. http://ceres-solver.org (accessed 08.05.2017).

Arie-Nachimson M, Kovalsky S Z, Kemelmacher-Shlizerman I, Singer A and Basri R (2012) Global motion estimation from point matches. *In Proceedings of the IEEE Conf. on 3D Vision*: 81-88.

Arrigoni F, Fusiello A, Rossi B, (2016) Camera motion from group synchronization. *In: Proceedings of the IEEE International conf. on 3D Vision*: 546-555.

Arya S, Mount D M, Netanyahu N S, Silverman R, Wu A Y (1998) An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *ACM*, 45(6): 891-923.

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

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J, et al, 2010. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3(1): 1–122.

Bourmaud G, Megret R, Giremus A, Berthoumieu Y (2014) Global motion estimation from relative measurements using iterated extended Kalman filter on matrix Lie groups. *In: Proceedings of the IEEE Conf. on Image Processing (ICIP)*: 3362- 3366.

Brand M, Antone M and Teller S (2004) Spectral solution of large scale extrinsic camera calibration as a graph embedding problem. *In: Proceedings of the European Conf. on Computer Vision (ECCV)*: 262-273.

Chatterjee A, Govindu V M (2013) Efficient and robust large-scale rotation averaging. *In: Proceedings of the IEEE International Conf. on Computer Vision (ICCV): 521–528.*

Chen L, Rottensteiner F, Heipke C (2016) Invariant descriptor learning using a Siamese convolutional neural network. *In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 3 (3):11-18.

Chen L, Rottensteiner F, Heipke C (2020) Deep Learning based Feature Matching and its Application in Image orientation. *In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. V-2-2020: 25–33.

Cohen A, Zach C, Sinha S N and Pollefeys M (2012) Discovering and exploiting 3d symmetries in structure from motion. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition(CVPR)*: 1514-1521.

Crandall D, Owens A, Snavely N, Huttenlocher D (2011) Discrete-continuous optimization for large scale structure from motion. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition(CVPR)*: 3001- 3008.

Cui H N, Gao X, Shen S H, Hu Z Y (2017) HSfM: Hybrid structure-from-motion. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition(CVPR)*.

Cui H N, Shen S H, Gao W, Liu H M, Wang Z H (2019) Efficient and robust large-scale structure-from-motion via track selection and camera prioritization. *ISPRS Journal of Photogrammetry & Remote Sensing*, 156: 202-214.

Cui Z, Jiang N, Tang C and Tan P (2015) Linear global translation estimation with feature tracks. *In Proceedings of British Machine Vision Conference.*

Cui Z, Tan P (2015) Global Structure-from-motion by Similarity Averaging. *In: Proceedings of the IEEE International Conf. on Computer Vision (ICCV)*: 864-872.

Duda R O and Hart P E (1973) Pattern Classification and Scene Analysis. *John Wiley and Sons.*

Farenzena M, Fusiello A, and Gherardi R (2009) Structureand-motion pipeline on a hierarchical cluster tree. *In: Proceedings of the IEEE International Conf. on Computer Vision (ICCV) Workshop*: 1489-1496.

Faugeras O (1992) What can be seen in three dimensions with an uncalibrated stereo rig? *In: Proceedings of the European Conference on Computer Vision (ECCV)*:563-578.

Fischler M and Bolles R (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communication of the Association for Computing Machinery (CACM)*, 24(6): 381–395.

Fitzgibbon A W, Zisserman A (1998) Automatic camera recovery for closed or open image sequences. *In: Proceedings of the European Conference on Computer Vision (ECCV)*: 311- 326.

Förstner, W (1986) A feature based correspondence algorithm for image matching. *International Archieves of Photogrammetry and Remote Sensing*, 26 (3): 150-166.

Förstner W, Wrobel B (2016) Photogrammetric Computer Vision: Statistics, Geometry, Orientation and Reconstruction. Springer.

Frahm J M, Fitegeorgel P, Gallup D, Johnson T, Raguram R, Wu C, et al. (2010) Building Rome on a cloudless day. *In: Proceedings of the European Conf. on Computer Vision (ECCV):* 368–381.

Geifman A, Kasten Y, Galun M, Basri R (2020) Averaging Essential and Fundamental Matrices in Collinear Camera Settings. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*

Gherardi R, Farenzena M, Fusiello A (2010) Improving the efficiency of hierarchical structure-and-motion. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR):* 1594-1600.

Gilmore R, (1974) Baker-campbell-hausdorff formulas. *Journal of Mathematical Physics*, 15(12): 2090-2092.

Govindu V M (2001) Combining two-view constraints for motion estimation. *In Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*: 218–225.

Govindu V M (2006) Robustness in motion averaging. In: Computer Vision – ACCV: 457-466.

Goldstein T, Hand P, Lee C, et al. (2016) Shapefit and shapekick for robust, scalable structure from motion. *In: Proceedings of the European Conf. on Computer Vision*: 289-304.

Harris C, Pike J (1988) 3d positional integration from image sequences. *In: Proceeding of Alvey Vision Conf:* 87–90.

Harris C, Stephens M (1988) A combined corner and edge detector. *In: Taylor CJ(ed) Proceedings of the Alvey Vision conference*. 15: 23.1-23.6.

Hartley R (1992) Estimation of relative camera positions for uncalibrated cameras. *In: Proceedings of the European Conference on Computer Vision (ECCV)*: 579-587.

Hartley R (1997) Lines and points in three views and the trifocal tensor. *International Jouranl of Computer Vision (IJCV)*, 22(2): 125-140.

Hartley R, Zisserman A (2003) Multiple View Geometry in Computer Vision. 2d ed. *Cambridge, UK: Cambridge University Press*.

Hartley R, Aftab K, Trumpf J (2011) L1 rotation averaging using the Weiszfeld algorithm. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR):* 3041–3048.

Hartley R, Trumpf J, Dai Y, Li H (2013) Rotation averaging. *International Journal of Computer Vision*, 103 (3): 267–305.

Hartmann W, Havlena M, Schindler K (2016) Recent developments in large-scale tie-point matching. *ISPRS Journal of Photogrammetry & Remote Sensing*, 115: 47-62.

Havlena M, Torii A, Knopp J, Pajdla T (2009) Randomized structure from motion based on atomic 3D models from camera triplets. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*: 2874-2881.

Havlena M, Schindler K (2014) VocMatch: efficient multiview correspondence for structure from motion. *In: Proceedings of the European Conf. on Computer Vision (ECCV).*

Heinly J, Dunn E and Frahm J M (2014) Correcting for duplicate scene structure in sparse 3D reconstruction. *In Proceedings of the European Conf. on Computer Vision (ECCV)*: 780-795.

Heinly J, Schonberger J L, Dunn E, Frahm J M (2015) Reconstructing the world* in six days. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*: 3287–3295

He K M, Zhang X Y, Ren S Q, Sun J (2016) Deep Residual Learning for Image Recognition. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*

Horn B K, Hilden H M, Negahdaripour S (1988) Close-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America A*, 5(7): 1127-1135.

Jiang N, Tan P and Cheong L F (2012) Seeing double without confusion: Structure-from-motion in highly ambiguous scenes. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition(CVPR)*: 1258-1465

Jiang N, Cui Z, Tan P (2013) A global linear method for camera pose registration. In: *Proceedings of the IEEE International Conf. on Computer Vision (ICCV): 481–488.*

Jiang S, Jiang W (2020) Efficient match pair selection for oblique UAV images based on adaptive vocabulary tree. *ISPRS Journal of Photogrammetry & Remote Sensing,* 161*:* 61-75.

Kasten Y, Geifman A, Galun M, Basri R (2019a). GPSfM: Global Projective SFM Using Algebraic Constraints on Multi-View Fundamental Matrices. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*

Kasten Y, Geifman A, Galun M, Basri R (2019b) Algebraic Characterization of Essential Matrices and Their Averaging in Multiview Settings. *In: Proceedings of the IEEE International Conf. on Computer Vision (ICCV).*

Krarup T, Juhl J, Kubik K (1980) Gotterdammerung over least squares adjustment. *International Archieves of Photogrammetry*, 23: 369-378.

Kraus K (1997) *Photogrammetry: Volume 2, advanced Methods and Appilcations*, volume 2. Duemmler Verlag, 4edition.

Krizhevsky A, Sutskever I and Hinton G E (2012) Imagenet classification with deep convolutional neural networks. *In: Proceedings of Advances in neural information processing systems*: 1097–1105.

Lowe D (2004) Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60 (2):91-110.

Longuet-Higgins HC (1981) A computer algorithm for reconstructinga scene from two projections. *Nature,* 293 (5828): 133–135.

Luhmann T, Robson S, Kyle S, Boehm J (2014) Close-Range photogrammetry and 3D Imaging. *De Gruyter*.

Luxburg V U (2007) A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395-416.

Martinec D and Pajdla T (2007) Robust rotation and translation estimation in multiview reconstruction. *In Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition:* 1– 8.

Mayer H (2014) Efficient hierarchical triplet merging for camera pose estimation. *In: Proceedings of German Conf. on Pattern Recognition*: 99–409.

Mayer H (2019) RPBA—Robust Parallel Bundle Adjustment Based on Covariance Information. *arXiv preprint arXiv:1910.08138*.

Michelini M, Mayer H (2020) Structure from motion for complex image sets. *ISPRS Journal of Photogrammetry & Remote Sensing, 166:* 140-152.

Moravec H (1980) Obstacle avoidance and navigation in the real world by a seeing robot rover. Ph.D. dissertation, Stanford Univ., Stanford, CA.

Moulon P, Monasse P (2012) Unordered feature tracking made fast and easy. *In: Proceedings of European Conference on Visual Media Production, CVMP*.

Moulon P, Monasse P, Marlet R (2013) Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion. *In Proceedings of the IEEE International Conf. on Computer Vision (ICCV):* 3248–3255.

Muja M, Lowe D G (2009) Fast approximate nearest neighbors with automatic algorithm configuration. *In: Proceeding of the International Conference on Computer Vision Theory and Applications (VISAPP)*: 331–340.

Muja M, Lowe D G (2012) Fast matching of binary features. *In: Proceeding of the IEEE Conf. on Computer and Robot Vision (CVR)*: 404-410.

Muja M, Lowe D G (2014) Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 3 (11): 2227–2240.

Mur-Artal R, Montiel J M M, Tardos J D (2015) ORB-SLAM: A Versatile and Accurate Monocular SLAM Symtem. *IEEE Transactions on Robotics*, 31 (5) : 1174-1163.

Nistér D (2000) Reconstruction from uncalibrated sequences with a hierarchy of trifocal tensors. *In: Proceedings of the European Conference on Computer Vision (ECCV)*: 649-663.

Nistér D (2004) An efficient solution to the five-point relative poseproblem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 (6): 756–770.

Nistér D, Stewenius H (2006) Scalable recognition with a vocabulary tree. *In: Proceedings of the IEEE International Conf. on Computer Vision and Pattern Recognition (CVPR)*: 2161–2168.

Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42 (3): 145–175.

Olsson C, Enqvist O (2011) Stable Structure from Motionfor Unordered Image Collections. In: *Proceeding of Scandinavian Conference on Image Analysis*. 6688: 524-535.

Özdemir E, Toschi I, and Remondino F (2019) A multi-purpose benchmark for photogrammetric urban 3D reconstruction in a controlled environment. *In: ISPRS Archives of Photogrammetry Remote Sensing and Spatial Information Science*, Vol. XLII-1/W2: 53–60.

Özyesil O, Singer A, Basri R (2015) Stable camera motion estimation using convex programming. *SIAM Journal of Image Science*, 8 (2): 1220–1262.

Pollefeys M, Van Gool L, Vergauwen M, Verbiest F, Cornelis K, Tops J, Koch R (2004) Visual modelling with a hand-held camera. *International Journal of Computer Vision (IJCV)*, 59 (3): 207-232.

Reich M, Heipke C (2015) Global rotation estimation using weighted iterative lie algebraic averaging. *In: ISPRS Annals of Photogrammetry Remote Sensing and Spatial Information Science*, II-3/W5: 443–449.

Reich M, Heipke C (2016) Convex image orientation from relativeorientations. *In: ISPRS Annals of Photogrammetry Remote Sensing and Spatial Information Science*, III-3: 107-114.

Reich M, Yang M Y, Heipke C (2017) Global robust image rotation from combined weighted averaging. *ISPRS Journal of Photogrammetry & Remote Sensing*, 127: 89-101.

Remondino F, EI-Hakim S (2006) Image-based 3D Modelling: A Review. The Photogrammetric Record, 21 (115): 269-291.

Remondino F, Gaiani M, Apollonio F, Ballabeni A, Ballabeni M, Moribito D (2016) 3D Documentation of 40 Kilometers of Historical Porticoes-the Challenge. In: *ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. XLI-B5: 711-718.

Ressl C (2000) An Introduction To The Relative Orientation Using The Trifocal Tensor. In*: International Archives of Photogrammetry and Remote Sensing*, 33 (B3): 769-776.

Roberts R, Sinha S N, Szeliski R and Steedly D (2011) Structure from motion for scenes with large duplicate structures. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*

Robinson J T (1984) k-d-tree splitting algorithm. *IBM Technical Disclosure Bulletin,* 27(5): 2974–2977.

Schmid H H (1958) Eine allgemeine analytische Lösung für die Aufgabe der Photogrammetrie. *Bildmessung und Luftbildwesen*, 26/27: 103-113.

Schönberger J L, Berg A C, Frahm J M (2015a) PAIGE: Pairwise image geometry encoding for improved efficiency in structure-from-motion. *In: Proceedings of the IEEE International Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Schönberger J L, Berg A C, Frahm J M (2015b) Efficient two-view geometry classification *In: Proceedings of German Conf. on Pattern Recognition*: 53–64.

Schönberger J L, Frahm J M (2016) Structure-from-Motion Revisited. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*

Shen T, Zhu S, Fang T, Zhang R and Quan L (2016) Graph based consistent matching for structure-from-motion. *In: Proceedings of the European Conference on Computer Vision (ECCV)*: 139–155.

Silpa-Anan C, Hartley R (2008) Optimised KD-trees for fast image descriptor matching. *In: Proceedings of the IEEE conf. on Computer Vision and Pattern Recognition (CVPR)*.

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. [Online]. Available: https://arxiv.org/abs/1409.1556

Sivic J, Zisserman A (2003) Video Google. A Text Retrieval Approach to Object Matching in Videos. *In Proceedings of the IEEE International Conf. on Computer Vision (ICCV)*.

Snavely N, Seitz SM, Szeliski R (2006) Photo Tourism: Exploring Photo Collection in 3D. In: *ACM Transactions on Graphics*. 25(3): 835-846.

Stewenius H, Engels C, Nistér D (2006) Recent developments on direct relative orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60 (4) :284-294.

Triggs B, McLauchlan P, Hartley R, Fitzgibbon A (2000) Bundle adjustment-a modern synthesis. In: *Vision Algorithms: Theory and Practice*, LNCS. Springer Verlag, 1883: 298-375.

Trzcinski T, Christoudias M, Lepetit V (2015) Learning image descriptors with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 37 (3): 597-610

Toldo R, Gherardi R, Farenzena M, Fusiello A (2015) Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision & Image Understanding*, 140:127-143.

Vedaldi A, Fulkerson B (2008). VLFeat: an open and portable library of computer vision algorithms. http://www.vlfeat.org

Wan J, Yilmaz A, Yan L (2018) DCF-BoW: Build Match Graph Using Bag of Deep Convolutional Features for Structure From Motion. *IEEE Geoscience and Remote Sensing letters*, 15 (12):1847-1851.

Wang F, Nayak A, Agrawal Y and Shilkrot R (2018a) Hierarchical Image Link Selection Scheme for Duplicate Structure Disambiguation. *In: Proceedings of British Machine Vision Conference (BMVC)*.

Wang X, Zhan Z Q, Heipke C (2017) An efficient method to detect mutual overlap of a large set of unordered images for structure-from. *In: ISPRS Annals of Photogrammetry Remote Sensing and Spatial Information Science*, IV-1/W1: 191-198.

Wang X, Rottensteiner F, Heipke, C (2018b) Robust image orientation based on relative rotations and tie points. *ISPRS Ann. Photogram., Rem. Sens. Spatial Inf. Sci.* IV-2: 295–302.

Wang X, Rottensteiner F, Heipke, C (2019a) Robust Structure from Motion based on relative rotations and tie points. *PE&RS*., 5: 347-359.

Wang X, Rottensteiner F, Heipke, C (2019b). Structure from motion for ordered and unordered image sets based on random k-d forests and global pose estimation. *ISPRS Journal of Photogrammetry & Remote Sensing*, 147: 19-41.

Wang X, Xiao T, Gruber M, Heipke C (2019c) Robustifying relative orientations with respect to repetitive structures and very short baselines for global SfM. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).*

Wang X, Heipke C (2020) An Improved Method of Refining relative orientation in Global Structure from Motion with a Focus on Repetitive Structure and Very Short Baselines. *PE&RS*., 5: 299-315.

Wang X, Xiao T, Kasten Y (2020) A Hybrid Global Image Orientation method for simultaneously Estimating Global Rotations and Global Translations. *ISPRS Ann. Photogram., Rem. Sens. Spatial Inf. Sci.* V-2-2020: 95-104.

Wang X, Xiao T, Kasten Y (2021) A Hybrid Global Structure from Motion method for synchronously Estimating Global Rotations and Global Translations. *ISPRS Journal of Photogrammetry & Remote Sensing*, 174: 35-55.

Weiszfeld E (1937) Sur le point pour lequel la somme des distance de n points donnes est minimum. *Tohoku Mathematical Journal*, 43 (2): 355-386.

Weiszfeld E, Plastria F (2009) On the point for which the sum of the distances to n given points is minimum. *Annals of Operations Research*, 167 (1): 7- 41.

Wilson K, Snavely N (2014) Robust Global Translations with 1DSfM. *In: Proceeding of the European Conf. on Computer Vision (ECCV) (pp. 61-75). Springer*

Wilson K, Bindel D, Snavely N (2016) When is Rotations Averaging Hard? *In: Proceeding of the European Conf. on Computer Vision (ECCV) (pp. 255-270). Springer*

Wu C, Agarwal S, Curless B, Seitz SM (2011) Multicore bundle adjustment. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*: 3057-3064.

Wu C (2013) Towards Linear-Time Incremental Structure from Motion. *In: Proceedings of the IEEE Conf. on 3dtv*: 127-134.

Zach C, Irschara A and Bischof H (2008) What can missing correspondences tell us about 3d structure and motion? *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*

Zach C, Klopschitz M and Pollefeys M (2010) Disambiguating visual relations using loop constraints. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*: 1426-1433.

Zhan Z Q (2006) Research on Camera Calibration Based on Completely Flat Liquid Crystal Display. PhD thesis, Wuhan University.

Zhan Z Q, Wang X, Wei ML (2015) Fast method of constructing image correlations to build a free network based on image multivocabulary trees. *Journal of Electronic Imaging*, 24 (3): 033029.

Zhan Z Q, Zhou G F, Yang X (2019) A Method of Hierarchical Image Retrieval for Real-Time Photogrammetry Based On Multiple Features. *IEEE Access*. 8: 21524-21533.

Zhang Z Y (2000) A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (11): 1330–1334.

Zhang R Z, Zhu S Y, Fang T (2017) Distributed Very Large Scale Bundle Adjustment by Global Camera Consensus. *In Proceedings of the IEEE International Conf. on Computer Vision (ICCV).*

Zhuang, B. B., Cheong, L. F., Lee, G. H. (2018) Baseline Desensitizing in Translation Averaging. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Curriculum Vitae

Personal Information

Xin Wang

18/12/1989 at HuangShi, China VR

Work Experience

10/2016 – 05/2021	Institute of Photogrammetry and GeoInformation (IPI) Leibniz University Hannover *Doctoral candidate*
01/2019 – 03/2019	Vexcel GmbH Graz, Austria *Secondment*

Education

09/2013 – 06/2016	School of Geodesy and Geomatics Wuhan University, Hubei, VR China *Master of Engineering*
09/2009 – 06/2013	School of Geodesy and Geomatics Wuhan University, Hubei, VR China *Bachelor of Engineering*
09/2006 – 06/2009	HuangShi No.2 Middle School, Hubei, VR China *Senior high school student*

Acknowledgements

On the occasion when my PhD studying is nearly finished, I would like to express my deepest thanks to all of the companions and supporters who have been with me during this PhD journey.

My first and deepest appreciation will go to myself. I thank myself for being kind of person who always has a very clear goal and tries to achieve that goal without any hesitation. Although the road to PhD is always full of challenges, constant efforts are quietly paid by myself. Doing research is actually sometimes very boring, but, it is not a problem for me anymore.

Then, I would also like to thank my supervisor Prof. Dr.-Ing. Habil. Christian Heipke. Thanks for offering me a chance to study at Institute of Photogrammetry and GeoInformation (IPI) in Hannover. During these four and half years, I learned a lot from you. Apart from the professional knowledge, the way of thinking and describing a problem from your side really influences me, which I believe will be life-long benefit to me. In addition, I also appreciate your concern on my life in Germany.

I would like to thank Prof. Franz Rottensteiner for all these discussions on my publications, especially in my early period publications, these discussions help me a lot in my later papers. Also, I would like to thank Prof. Steffen Schön and Prof. Helmut Mayer for acting as referee and Prof. Philipp Otto for chairing the examination committee. Furthermore, I would like also to thank the EU Volta project and Dr. Michael Gruber for supporting me as a visiting student at Vexcel Imaging company, there I had an unforgettable two months. And, I would like to thank the Chinese Scholarship Council (CSC) for financially funding my PhD studying in Hannover from 2016 to 2020.

I thank all the colleagues at IPI for a very friendly and happy studying time. In particular, I really appreciate the efforts that Max Mehltretter had spent on proof-reading my thesis, and Christian Kruse for being a so nice officemate with me for four years. And, I would like to say thank you to my Chinese friends (Dr.-Ing. Lin Chen, Jing Yang, Chun Yang, Yu Feng, Zhe Zeng etc.), thanks for your companion and the happiness from you.

Lastly, I would like to thank my family, my parent and my sister, they are always very proud of me, and give me full and pure support. Thanks to their encouragement, this thesis can be successfully finished.

(Eine vollständige Liste der Wiss. Arb. ist beim Geodätischen Institut, Nienburger Str. 1, 30167 Hannover erhältlich.)

Nr. 346 SCHLICHTING, Alexander: Fahrzeuglokalisierung durch Automotive Laserscanner unter Verwendung statischer Merkmale (Diss. 2018)
Nr. 347 RÖTH, Oliver Extraktion von hochgenauer Fahrspurgeometrie und -topologie auf der Basis von Fahrzeugtrajektorien und Umgebungsinformationen (Diss. 2018)
Nr. 348 NEELMEIJER, Julia: Observing Inter- and Intra-Annual Glacier Changes and Lake Loading Effects from Synthetic Aperture Radar Remote Sensing (Diss. 2018)
Nr. 349 HOBERG, Thorsten: Conditional Random Fields zur Klassifikation multitemporaler Fernerkundungs daten unterschiedlicher Auflösung (Diss. 2018)
Nr. 350 SCHILLING, Manuel: Kombination von klassischen Gravimetern mit Quantensensoren (Diss. 2019)
Nr. 351 MILLER, Dominik: Seismic noise analysis and isolation concepts for the ALPS II experiment at DESY (Diss. 2019)
Nr. 352 ALI, Bashar: Optimierte Verteilung von Standorten der Schulen unter dem Einfluss des demografischen Wandels am Beispiel Grundschulen (Diss. 2019)
Nr. 353 ZHAO, Xin: Terrestrial Laser Scanning Data Analysis for Deformation Monitoring (Diss. 2019)
Nr. 354 HAGHIGHI, Mahmud Haghshenas: Local and Large Scale InSAR Measurement of Ground Surface Deformation (Diss. 2019)
Nr. 355 BUREICK, Johannes: Robuste Approximation von Laserscan-Profilen mit B-Spline-Kurven (Diss. 2020)
Nr. 356 BLOTT, Gregor: Multi-View Person Re-Identification (Diss. 2020)
Nr. 357 MAAS, Alina Elisabeth: Klassifikation multitemporaler Fernerkundungsdaten unter Verwendung fehlerbehafteter topographischer Daten (Diss. 2020)
Nr. 358 NGUYEN, Uyen: 3D Pedestrian Tracking Using Neighbourhood Constraints (Diss. 2020)
Nr. 359 KIELER, Birgit: Schema-Matching in räumlichen Datensätzen durch Zuordnung von Objektinstanzen (Diss. 2020)
Nr. 360 PAUL, Andreas: Domänenadaption zur Klassifikation von Luftbildern (Diss. 2020)
Nr. 361 UNGER, Jakob: Integrated Estimation of UAV Image Orientation with a Generalised Building Model (Diss. 2020)
Nr. 362 COENEN, Max: Probabilistic Pose Estimation and 3D Reconstruction of Vehicles from Stereo Images (Diss. 2020)
Nr. 363 GARCIA FERNANDEZ, Nicolas: Simulation Framework for Collaborative Navigation: Development - Analysis – Optimization (Diss. 2020)
Nr. 364 VOGEL, Sören: Kalman Filtering with State Constraints Applied to Multi-sensor Systems and Georeferencing (Diss. 2020)
Nr. 365 BOSTELMANN, Jonas: Systematische Bündelausgleichung großer photogrammetrischer Blöcke einer Zeilenkamera am Beispiel der HRSC-Daten (Diss. 2020)
Nr. 366 OMIDALIZARANDI, Mohammad: Robust Deformation Monitoring of Bridge Structures Using MEMS Accelero-meters and Image-Assisted Total Stations (Diss. 2020)
Nr. 367 ALKHATIB, Hamza: Fortgeschrittene Methoden und Algorithmen für die computergestützte geodätische Datenanalyse (Habil. 2020)
Nr. 368 DARUGNA, Francesco: Improving Smartphone-Based GNSS Positioning Using State Space Augmentation Techniques (Diss. 2021)
Nr. 369 CHEN, Lin: Deep learning for feature based image matching (Diss. 2021)
Nr. 370 DBOUK, Hani: Alternative Integrity Measures Based on Interval Analysis and Set Theory (Diss. 2021)
Nr. 371 CHENG, Hao: Deep Learning of User Behavior in Shared Spaces (Diss. 2021)
Nr. 372 MUNDT Reinhard Walter: Schätzung von Boden- und Gebäudewertanteilen aus Kaufpreisen bebauter Grundstücke (Diss. 2021)
Nr. 373 WANG, Xin: Robust and Fast Global Image Orientation (Diss. 2021)

Die Arbeiten werden im Rahmen des wissenschaftlichen Schriftenaustausches verteilt und sind nicht im Buchhandel erhältlich. Der Erwerb ist zu einem Stückpreis von € 25,00 bei den herausgebenden Instituten möglich