DGK Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 863

Stefan Cavegn

Integrated Georeferencing for Precise Depth Map Generation Exploiting Multi-Camera Image Sequences from Mobile Mapping

München 2020

Verlag der Bayerischen Akademie der Wissenschaften

ISSN 0065-5325 | ISBN 978-3-7696-5275-8

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie der Universität Stuttgart
https://dx.doi.org/10.18419/opus-11210, Stuttgart 2020

Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 863

Integrated Georeferencing for Precise Depth Map Generation Exploiting Multi-Camera Image Sequences from Mobile Mapping

Von der Fakultät für Luft- und Raumfahrttechnik und Geodäsie

der Universität Stuttgart

zur Erlangung des Grades

Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation

Vorgelegt von

M.Sc. Stefan Cavegn

Geboren am 14.09.1985 in Ilanz/Glion, Schweiz

München 2020

Verlag der Bayerischen Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5275-8

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie der Universität Stuttgart
https://dx.doi.org/10.18419/opus-11210, Stuttgart 2020

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.-Ing. Uwe Sörgel

Referent: apl. Prof. Dr.-Ing. Norbert Haala

Korreferenten: Prof. Dr.-Ing. habil. Dr. h.c. Volker Schwieger
Prof. Dr. Stephan Nebiker

Tag der mündlichen Prüfung: 18.09.2020

ISSN 0065-5325 ISBN 978-3-7696-5275-8

Contents

Acronyms

CCD	Charge-Coupled Device
CNN	Convolutional Neural Network
COLMAP	Open-source incremental structure-from-motion tool
CP	Check Point
DG	Direct Georeferencing
DIM	Dense Image Matching
DoF	Degree of Freedom
DSM	Digital Surface Model
EOPs	Exterior Orientation Parameters
FHD	Full High Definition (Resolution of 1920 × 1080 pixels)
FHNW	University of Applied Sciences and Arts Northwestern Switzerland
FOV	Field of View
fps	frames per second
GCP	Ground Control Point
GNSS	Global Navigation Satellite Systems
GSD	Ground Sampling Distance
IG	Integrated Georeferencing
IGEO	Institute of Geomatics
IMU	Inertial Measurement Unit
INS	Inertial Navigation System
IOPs	Interior Orientation Parameters
LA	Lever Arm
Lidar	Light detection and ranging
LN02	Swiss vertical reference frame
LV03/LV95	Swiss horizontal reference frames
MA	MisAlignment

MEMS	Micro-Electro-Mechanical Systems
MMS	Mobile Mapping System
MP	MegaPixel
MVS	Multi-View Stereo
RANSAC	Random Sample Consensus
RGB	Image with Red, Green and Blue channels
RGB-D	Image with Red, Green, Blue and Depth channels
RMSE	Root Mean Square Error
ROC	Relative Orientation Constraint
ROPs	Relative Orientation Parameters
SD	Standard Deviation
SfM	Structure from Motion
SGM	Semi-Global Matching
SIFT	Scale Invariant Feature Transform
SLAM	Simultaneous Localization and Mapping
SURE	Software of nFrames for photogrammetric SUrface REconstruction
TLS	Terrestrial Laser Scanning
UAV	Unmanned Aerial Vehicle
WGS84	World Geodetic reference System 1984

Abstract

Image-based mobile mapping systems featuring multi-camera configurations allow for efficient geospatial data acquisition in both outdoor and indoor environments. We aim at accurate geospatial 3D image spaces consisting of collections of georeferenced multi-view RGB-D imagery, which may serve as basis for 3D street view services. In order to obtain high-quality depth maps, dense image matching exploiting multi-view image sequences captured with high redundancy needs to be performed. Since this process is entirely dependent on accurate image orientations, we mainly focus on pose estimation of multi-camera systems within this thesis. Nonetheless, we also present methods and investigations to obtain accurate, reliable and complete 3D scene representations based on multi-stereo mobile mapping sequences.

Conventional image orientation approaches such as direct georeferencing enable absolute accuracies at the centimeter level in open areas with good GNSS coverage. However, GNSS conditions of street-based mobile mapping in urban canyons are often deteriorated by multipath effects and by shading of the signals caused by vegetation and large multi-story buildings. Moreover, indoor spaces do not even allow for any GNSS signals. Hence, we propose a powerful and versatile image orientation procedure that is able to cope with these issues encountered in challenging urban environments.

Our integrated georeferencing approach extends the powerful structure-from-motion pipeline COLMAP with georeferencing capabilities. It assumes initial camera poses with sub-meter accuracy, which allow for direct triangulation of the complete scene. Such a global approach is much more efficient than an incremental structure-from-motion procedure. Furthermore, an initial image orientation solution already facilitates to georeference in a geodetic reference frame. Nevertheless, accuracies at the centimeter level can only be achieved by incorporation of ground control points. In order to obtain sub-pixel accurate relative orientations, strong tie point connections for the highly redundant multi-view image sequences are required. However, hardly overlapping fields of view, strongly varying views and weakly textured surfaces aggravate image feature matching. Hence, constraining relative orientation parameters among cameras is crucial for accurate, robust and efficient image orientation. Apart from supporting fixed multi-camera rigs, our integrated georeferencing approach that uses bundle adjustment allows for self-calibration of all relative orientation parameters or just single components.

We extensively evaluated our integrated georeferencing procedure using six challenging real-world datasets in order to demonstrate its accuracy, robustness, efficiency and versatility. Four datasets were captured outdoors, one by a rail-based and three by different street-based multi-stereo camera systems. A portable mobile mapping system featuring a multi-head panorama camera collected two datasets in an indoor environment. Employing relative orientation constraints and ground control points within these indoor spaces resulted in absolute 3D accuracies of ca. 2 cm, and precisions at the millimeter level for relative 3D measurements. Depending on the use case, absolute 3D accuracy values for outdoor environments are slightly larger and amount to a few centimeters. However, determining 3D reference coordinates is a costly task. Not relying on any ground control points led to horizontal accuracies of ca. 5 cm for a scenario featuring some loops, while dropping down to a few decimeters for an extended junction area. Since the height component is even more dependent on prior camera poses from direct georeferencing, these 2D accuracies significantly decreased for the 3D case. However, incorporating just one ground control point facilitates the elimination of systematic effects, which results in 3D accuracies within the sub-decimeter range. Nevertheless, at least one additional check point is recommended in order to ensure a reliable solution.

Once consistent and sub-pixel accurate relative poses of spatially adjacent images are available, in-sequence dense image matching can be performed. Aiming at precise and dense depth map generation, we evaluated several image matching configurations. Standard single stereo matching led to high accuracies, which could not significantly be improved by in-sequence matching. However, the image redundancy provided by additional epochs resulted in more complete and reliable depth maps.

Kurzfassung

Bildbasierte Mobile Mapping Systeme, die mit Mehr-Kamera-Konfigurationen ausgestattet sind, ermöglichen eine effiziente räumliche Datenerfassung im Aussen- wie auch im Innenraum. Ziel ist die Erzeugung von georeferenzierten Multi-View RGB-D Bildern als Grundlage für 3D-Bildräume bzw. 3D Street View Services. Um dafür hochwertige Tiefenkarten zu erhalten, ist eine dichte Bildzuordnung der mit hoher Redundanz erfassten Multi-View Bildsequenzen erforderlich. Da dieser Prozess aber gänzlich von genauen Bildorientierungen abhängig ist, wird in dieser Arbeit hauptsächlich auf die Posenbestimmung von Mehr-Kamera-Systemen fokussiert. Es werden jedoch auch Methoden und Untersuchungen präsentiert, welche basierend auf Multi-Stereo Mobile Mapping Sequenzen der Generierung von genauen, zuverlässigen und vollständigen 3D-Szenenrepräsentationen dienen.

Konventionelle Bildorientierungsansätze, wie beispielsweise die direkte Georeferenzierung, ermöglichen absolute Genauigkeiten im Zentimeterbereich in offenen Gebieten die eine gute GNSS-Abdeckung aufweisen. GNSS-Bedingungen für strassenbasiertes Mobile Mapping in urbanen Strassenschluchten werden aber oft durch grosse mehrstöckige Gebäude und Vegetation beeinträchtigt, was zu Mehrwegeffekten und Signalabschattungen führt. Zudem können in Innenräumen überhaupt keine GNSS-Signale empfangen werden. Somit wird ein leistungsfähiges und vielseitiges Bildorientierungsverfahren vorgeschlagen, welches diese Schwierigkeiten, die in anspruchsvollen urbanen Umgebungen anzutreffen sind, bewältigen kann.

Der entwickelte integrierte Georeferenzierungsansatz erweitert die leistungsfähige Structure-from-Motion Pipeline COLMAP mit Georeferenzierungsfunktionalität. Es werden initiale Kameraposen mit Sub-Meter-Genauigkeit vorausgesetzt, die eine direkte Triangulation der gesamten Szene ermöglichen. Solch ein globaler Ansatz ist um einiges effizienter als ein inkrementelles Structure-from-Motion Verfahren. Eine initiale Bildorientierungslösung lässt auch schon die Georeferenzierung in einem geodätischen Referenzrahmen zu. Genauigkeiten im Zentimeterbereich können aber nur unter Verwendung von Bodenpasspunkten erreicht werden. Um Sub-Pixel genaue relative Orientierungen zu erhalten, werden starke Verknüpfungspunktbeziehungen zwischen den hochredundanten Multi-View Bildsequenzen benötigt. Kaum überlappende Sichtfelder, sehr unterschiedliche Ansichten und schwach texturierte Oberflächen erschweren jedoch die Bildmerkmalszuordnung. Demzufolge ist die Einführung von Bedingungen für relative Orientierungsparameter zwischen Kameras entscheidend für eine genaue, robuste und effiziente Bildorientierung. Neben der Unterstützung von festen Mehr-Kamera-Anordnungen ermöglicht der integrierte Georeferenzierungsansatz mit Bündelausgleichung eine Selbstkalibrierung von allen relativen Orientierungsparametern oder auch nur von einzelnen Komponenten.

Das integrierte Georeferenzierungsverfahren wurde anhand von sechs anspruchsvollen Datensätzen umfassend evaluiert, um dessen Genauigkeit, Robustheit, Effizienz und Vielseitigkeit zu demonstrieren. Vier Datensätze wurden im Aussenraum erfasst, einer mit einem schienenbasierten und drei mit unterschiedlichen strassenbasierten Multi-Stereo Kamerasystemen. Ein portables Mobile Mapping System, das über eine Mehr-Kopf-Panoramakamera verfügt, ermöglichte die Erfassung von zwei Datensätzen in einem Innenbereich. Die Verwendung von relativen Orientierungsbedingungen und Bodenpasspunkten in diesen Innenräumen führte zu absoluten 3D-Genauigkeiten von ca. 2 cm, wie auch zu Präzisionswerten im Millimeterbereich für relative 3D-Messungen. Je nach Anwendungsfall wurden für den Aussenraum etwas grössere absolute 3D-Genauigkeitswerte im Zentimeterbereich ermittelt. Die Bestimmung von 3D-Referenzkoordinaten ist jedoch arbeitsaufwändig. Bei Nichtberücksichtigung von Bodenpasspunkten wurden horizontale Genauigkeiten von ca. 5 cm für ein Szenario mit einigen Schleifen erreicht, während

ein erweiterter Strassenkreuzungsbereich zu mehreren Dezimetern führte. Da die Höhenkomponente noch stärker von den initialen Kameraposen aus der direkten Georeferenzierung abhängig ist, verschlechterten diese 2D-Genauigkeiten signifikant für den 3D-Fall. Wird aber nur ein Bodenpasspunkt einbezogen, können systematische Effekte eliminiert werden, was zu 3D-Genauigkeiten im Sub-Dezimeter-Bereich führt. Um eine zuverlässige Lösung zu gewährleisten, wird jedoch empfohlen, zusätzlich mindestens einen Kontrollpunkt zu verwenden.

Sobald konsistente und Sub-Pixel genaue relative Posen von räumlich benachbarten Bildern vorliegen, kann eine sequenzbasierte dichte Bildzuordnung ausgeführt werden. Mehrere Bildzuordnungskonfigurationen wurden für die Generierung von möglichst präzisen und dichten Tiefenkarten evaluiert. Standard Einzel-Stereo Matching ergab hohe Genauigkeiten, die durch ein Matching in der Sequenz nicht signifikant verbessert werden konnten. Zusätzliche Epochen lieferten aber eine erhöhte Bildredundanz, die zu vollständigeren und zuverlässigeren Tiefenkarten führte.

Chapter 1

Introduction

1.1 Motivation - Urban 3D Data Capture from Mobile Mapping Image Sequences

Prevalent digitalization trends aiming at generating high-fidelity scene representations or digital twins demand 3D geoinformation of high quality. Large-scale areas can be mapped efficiently by airborne platforms, which typically deliver accuracies that correspond to 1 GSD. However, targeted geometric resolutions and thus accuracies for such scenarios typically amount to one decimeter, which is often not sufficient for 3D mapping of urban road environments. Moreover, occlusions caused by buildings and vegetation lead to incomplete data collection. In contrast, vehicle-based mobile mapping systems are ideally suited for efficient and accurate 3D data capture of traffic corridors, where a large part of the public technical infrastructure is located. Hence, resulting digital 3D realities can be exploited e.g. for infrastructure asset inventories and efficient road or rail infrastructure management, thus avoiding dangerous fieldwork.

First successful mobile mapping experiments led to the GPSVan (Novak, 1991) and the VISAT system (Schwarz et al., 1993), which were based on stereo camera systems. Nonetheless, for a considerable period, street-based mobile mapping systems usually featured lidar sensors as main components (Puente et al., 2013), and used cameras solely for point cloud colorization. Lidar delivers moderate densities and high precision. Since it is an active acquisition method, even homogeneous surfaces and difficult lighting conditions can be overcome, which is particularly beneficial in indoor environments. However, successful co-registration of images and laser scans poses a few challenges. Even if synchronization as well as calibration of offsets and rotations among laser scanners and cameras are managed precisely, moving objects such as pedestrians and vehicles can still lead to inconsistencies. Camera-based mobile mapping systems can prevent such issues, since both radiometry and derived 3D geometry are based on the same source. Furthermore, image-based measurements and scene interpretations are far more intuitive than measurements in 3D point clouds. They also require very little training, in contrast to measurements in 3D point clouds, which typically need expert skills. An extensive comparison between lidar and vision is given by Leberl et al. (2010).

Significant progress in imaging sensors with respect to both geometric and radiometric resolution, as well as advances in image processing methods, have made camera-based systems even more attractive for highly efficient and accurate 3D mapping in complex urban environments (Pollefeys et al., 2008; Gallup, 2011). Many powerful algorithms for dense 3D scene generation rely on stereo image matching. There are two main approaches for obtaining stereo images, either by establishing physical or virtual stereo bases. The first requires two cameras that are precisely synchronized and rigidly attached to a frame, while the latter uses images captured by the same camera at two different epochs. Precalibrated physical stereo bases are typically an order of magnitude more precise than virtual stereo baselines resulting from vehicle movement. Hence, we prefer multi-stereo camera configurations to meet our high accuracy requirements (Cavegn and Haala, 2016). In order to obtain a full 360° 3D coverage in urban areas with multi-story buildings and numerous superstructures, we developed a novel 360° panoramic stereo camera configuration

(Blaser et al., 2018). It uses two multi-head 360° panorama cameras, tilted forward and backward by 90° respectively, and offers large rigid stereo bases for all viewing directions. Such multi-view stereo configurations are well suited for generating accurate geospatial 3D image spaces consisting of collections of georeferenced multi-view RGB-D imagery (Nebiker et al., 2015). The underlying images combined with depth maps derived from dense image matching allow for a high-fidelity scene representation with an unparalleled level of detail. Furthermore, they can be exploited for tasks such as 3D monoplotting enabling a user to accurately determine 3D coordinates of features of interest simply by clicking on a location within the 2D images. However, depth maps are frequently generated by performing multi-view stereo matching using imagery captured at different epochs. In order to efficiently apply coplanarity constraints during dense stereo matching, sub-pixel accurate relative orientations of the image sequences are required.

In contrast to airborne nadir applications, where ground sampling distances remain roughly constant over the complete mapping area, vehicle-based mobile mapping images show large scale variations caused by different distances to mapping objects. Hence, using a common mobile mapping configuration, one pixel corresponds to 2-6 mm in object space for a typical measurement range of 4-14 m and it is 1 cm at 23 m (Cavegn and Haala, 2016). While infrastructure management applications often demand 3D measurement accuracies within the sub-decimeter range, urban modeling requires absolute accuracies at the centimeter level. As earlier studies showed, these requirements could already be met in open areas. Burkhard et al. (2012) obtained absolute 3D point measurement accuracies of 4-5 cm in average to good GNSS conditions using their stereovision mobile mapping system. The capability of the StreetMapper lidar mobile mapping system to produce dense 3D measurements at an accuracy level of 3 cm in good GNSS conditions was demonstrated by Haala et al. (2008). Nonetheless, GNSS conditions of land-based mobile mapping vehicles in urban environments are often deteriorated by multipath effects and by shading of the signals caused by trees and large multi-story buildings, which aggravate fulfilling the accuracy requirements by only performing direct georeferencing. Furthermore, distances between cameras and measured objects are typically a few meters, compared to several hundred meters for airborne applications. Therefore, the contribution of the GNSS positioning error to the overall error budget is much larger than the contribution of the error from the attitude determination. Moreover, Pollefeys et al. (2008) and Frahm et al. (2010) encountered trajectory discontinuities of ca. 10 cm in direct georeferencing during a vehicle stop of several seconds. Such inconsistencies can only be corrected by image-based georeferencing, which additionally allows for the elimination of trajectory offsets in the range of several decimeters. Hence, an integrated georeferencing approach using bundle adjustment and incorporating multi-view image sequences aiming at improving accuracy, robustness and efficiency needed to be developed.

1.2 Objectives - Georeferencing of Multi-View Imagery and 3D Computation in Urban Environments

Accurate image orientations are the foundation of all subsequent steps of a 3D mapping pipeline. This pipeline typically includes depth map computation, point cloud and mesh generation as well as texturing. Densely built-up urban environments with extended areas of poor GNSS coverage frequently limit direct georeferencing accuracies to several decimeters up to meters. Furthermore, indoor environments prevent the use of GNSS signals, so that image orientation solutions require alternative approaches such as simultaneous localization and mapping (SLAM). Both direct georeferencing and SLAM do usually not provide sub-pixel accurate relative poses of spatially adjacent images, which are needed for dense image matching. Hence, it was our main objective to develop a powerful and versatile integrated georeferencing approach for challenging outdoor and indoor environments. We targeted to significantly increase the initial georeferencing quality in terms of accuracy, robustness and efficiency by exploiting highly redundant multi-view image sequences captured by camera-based mobile mapping systems. The new georeferencing approach should address the following main issues:

Accuracy Direct georeferencing or SLAM typically delivers accuracies of several decimeters in challenging environments. This does neither meet our absolute accuracy requirements of a few centimeters nor provide image orientation accuracies at the sub-pixel level that is needed for satisfying dense image matching results. However, camera poses featuring accuracies within the meter range are well suited as initial solution for advanced integrated georeferencing.

Robustness Multi-camera systems pointing in various directions allow for large fields of view. However, if treated individually, vision-based pipelines are usually not able to successfully orient all images. Therefore, imagery captured at the same epoch have to be handled as one set.

Efficiency Determination of 3D coordinates of ground control points is a time-consuming and costly task. Hence, an integrated georeferencing approach relying on just a few or even no ground control points was desired. While real-time approaches are mandatory in the robotics industry, short processing times were not our main interest.

Versatility Our integrated georeferencing approach needs to be applicable universally. In addition to challenging urban road and rail environments, it should also perform well indoors where no GNSS signals are available and feature matching conditions are aggravated due to low-textured surfaces or repetitive patterns. Furthermore, supporting multi-stereo pinhole camera systems and multi-head panorama cameras as well as the corresponding perspective and fisheye camera models is crucial.

Our overall goal is the generation of high-quality depth maps that build a crucial part of georeferenced RGB-D imagery. Since they are entirely dependent on accurate image orientations, we mainly focus on integrated georeferencing of multi-camera systems. Nonetheless, methods and investigations to obtain accurate, reliable and complete 3D scene representations based on multi-stereo mobile mapping sequences are of interest as well. In order to achieve these aims, leveraging the high image redundancy is essential.

Accuracy Our intended applications based on depth maps demand absolute accuracies within the sub-decimeter range and relative accuracies at the centimeter level. Shortened measurement distances provide high accuracies, but require sufficient image capturing rates, i.e. collection of multiple images every two meters. Furthermore, adequate stereo bases constitute a key component for precise relative measurements.

Reliability We expect to obtain depth maps that feature a low amount of outliers. Redundancy allows that each surface element is seen by multiple views, which is the factor of success for each filtering technique.

Completeness Dense image matching allows to compute a 3D depth value for each single pixel of the area that is covered by at least two images. Difficulties arise due to vegetation, reflective or transparent surfaces such as glass façades or windows. Moving objects are challenging, but stereo matching enables a spatial and temporal coherence.

1.3 Main Contributions - Integrated Georeferencing Exploiting Relative Orientation Constraints

Mobile mapping systems featuring multi-view stereo or multi-head cameras capture a set of several images at each epoch. In order to successfully orient such multi-view image sequences, relative orientation constraints among cameras need to be exploited within an integrated georeferencing procedure. Such methods require a multitude of prior metric information. Hence, we assume the availability of the following:

- precisely calibrated interior orientation parameters (IOPs) of all cameras
- precisely calibrated relative orientation parameters (ROPs) among cameras
- calibrated lever arm (LA) and misalignment (MA) parameters by boresight alignment
- initial exterior orientation parameters (EOPs) from direct georeferencing or SLAM
- limited number of ground control points (GCPs) in a predefined geodetic reference frame

We extended a powerful structure-from-motion pipeline with *georeferencing capabilities*. While highly redundant multi-view image sequences lead to strong image connections and thus accurate relative orientations, prior camera poses and ground control points allow for georeferencing in a geodetic reference frame. Based on initial EOPs, we perform *spatial feature matching* by only considering image candidates that lie within a predefined maximum range and a specified field of view. This *search space reduction* leads to a significant speed-up of the process. A further efficiency increase was reached by modifying the existing incremental SfM pipeline into a *global approach*. Hence, prior camera poses allow for immediate triangulation of the complete scene, which does not suffer from a weak initial pair initialization. Furthermore, initial EOPs always enable a *weak datum computation*. Incorporation of ground control information shows the highest accuracy potential, but determination of reference data is costly.

Exploitation of constraints for the precisely calibrated offsets and rotations among respective cameras is one of our main contributions. Provided that all cameras are rigidly attached to a platform, *constraining ROPs* allows for more robust and accurate image orientation results. While stereo cameras have large overlapping fields of view, there is usually no overlapping area among individual stereo systems of a multi-stereo configuration. Furthermore, multi-head panorama cameras feature small overlapping areas. Therefore, sufficient tie point connections cannot be established among all images captured at the same point of time, which results in many non-oriented images. Multi-view image sequences captured in opposite driving directions can diminish this issue. However, well textured surfaces are still required for sufficient feature extraction and matching, which is especially not the case for indoor environments. Hence, *enforcing consistent ROPs for all epochs* is the key element. Apart from supporting fixed multi-camera rigs, our integrated georeferencing approach that performs bundle adjustment allows for *self-calibration* of all relative orientation parameters or just single components. There is no need for a calibration field on-site, since collecting multi-view images in opposite driving directions frequently suffices. If precise ROPs and initial EOPs are available, *employment of GCPs is optional*. While such cost-efficient processings result in well co-registered multi-view image sequences from multiple driving directions and campaigns, and thus lead to precise relative poses, absolute accuracies of several decimeters have to be expected.

We *comprehensively evaluated* our developed georeferencing approach in order to demonstrate its accuracy, robustness, efficiency and versatility. *Four datasets* were captured *outdoors* by either a rail-based or varying street-based multi-stereo mobile mapping systems. Furthermore, we collected *two datasets* in an *indoor environment* by a portable panoramic mobile mapping system. The fact that all scenarios feature real-world conditions is of high practical relevance.

Once consistent and sub-pixel accurate stereo image sequences are available, in-sequence dense image matching for precise depth map generation can be performed. Due to movement predominantly in camera viewing direction, a polar rectification method was required to fully exploit the redundancy provided by images captured at multiple epochs. We show that single stereo matching delivers high accuracies, whereas in-sequence matching leads to more reliable and complete 3D scene representations.

1.4 Outline

We perform image-based 3D reconstruction based on highly redundant multi-view image sequences captured by mobile mapping systems. Since we mainly focus on sophisticated computation of accurate camera poses, this thesis is structured in the following four main parts: related work, developed image orientation methods, evaluation of integrated georeferencing approach, evaluation of in-sequence dense image matching.

Chapter 2 provides a literature review comprising both image orientation and dense 3D scene generation approaches. First, general image orientation concepts derived from different research communities are presented and compared. Second, the focus is laid on pose estimation of multi-camera systems in challenging urban environments that do not allow for a sufficient number of GNSS signals. Afterwards, a survey covering dense image matching procedures for the computation of 3D urban representations is given.

Chapter 3 introduces our developed integrated georeferencing approach. It extends the powerful structure-from-motion pipeline COLMAP with georeferencing capabilities. Apart from incorporating initial camera poses from direct georeferencing or SLAM as well as ground control points, relative orientation constraints among cameras are exploited. Our procedure enables image orientation accuracies at the sub-pixel level, which are required for multi-view stereo matching of high-quality.

Chapter 4 comprehensively evaluates our integrated georeferencing approach based on six datasets featuring different environments and varying multi-camera configurations. Well textured road and rail scenes offer favorable conditions for establishing image feature correspondences. In contrast, indoor environments that primarily show repetitive structures and homogeneous surfaces are more challenging. In addition to employing prior camera poses, ground control points and relative orientation constraints, investigations include self-calibration of relative orientation parameters among cameras as well as no utilization of ground control points.

Chapter 5 presents an evaluation of in-sequence dense image matching. Varying image configurations also comprising multiple epochs are compared in order to generate dense and precise depth maps as well as accurate 3D point clouds. Furthermore, the multi-view stereo matching process can assess whether the image orientations provided by our integrated georeferencing approach meet the requirements of sub-pixel accuracy. We utilized the SURE software system with its implemented polar rectification approach that is able to handle large motion in camera viewing direction.

Chapter 6 concludes this thesis by summarizing the potential of our integrated georeferencing approach. Moreover, it is compared to state-of-the-art procedures and recommendations for multi-camera configurations are given. In addition, open issues and future work are discussed, while some recent learning-based approaches for both image orientation and dense 3D scene computation are presented.

Chapter 2

From Imagery to 3D Geometry

Image-based 3D reconstruction includes an image orientation and a 3D computation process. Hence, we first present and compare general image orientation concepts derived from different research communities, namely photogrammetry, computer vision and robotics. Afterwards, we focus on pose estimation of stereo and multi-view image sequences captured by mobile mapping systems. To this end, urban canyons with poor GNSS coverage and indoor spaces that do not allow for any GNSS signals are particularly challenging. Since direct georeferencing cannot deliver sufficient accuracies in such environments, an image-based georeferencing approach for pose estimation of multi-camera systems is required. Our survey further covers dense image matching procedures for subsequent computation of complex 3D geometry encountered in urban areas. While learning-based approaches for both image orientation and dense 3D scene generation are beyond the scope of this thesis, we present some of these recent methods in section 6.2.

2.1 Image Orientation Approaches for Diverse Scenarios

The goal of image orientation procedures is to determine 3D position and attitude information for all images at acquisition time. While traditionally tackled by the geodesy and photogrammetry communities, other communities have addressed this problem in the past decades as well. Due to emerging technologies such as autonomous driving and augmented reality, image orientation is still a hot topic. Photogrammetry mainly focuses on accuracy and reliability, while speed and robustness is more important in robotics. However, all image orientation approaches share the same important component that is bundle adjustment.

2.1.1 Photogrammetric Georeferencing Concepts

The photogrammetry community differentiates three image orientation procedures (see figure 2.1). Traditional **indirect georeferencing** relies on tie point connections established within overlapping aerial image regions as well as ground control points (GCPs). The 3D coordinates of these GCPs are previously determined accurately in a geodetic reference frame by tachymetry or GNSS. In order to comply with the 7 degrees of freedom of a 3D Helmert transformation, at least three reference points are needed. However, highly accurate and homogeneous results usually demand many more GCPs that are well distributed. Bundle block adjustment, also referred to as aerial triangulation, computes both exterior orientation parameters for each single image and 3D tie point coordinates. If desired, even interior orientation parameters can be self-calibrated. Linearized observation equations require initial values, which can be derived from flight mission planning or aircraft navigation data. Indirect georeferencing approaches were already established several decades ago, e.g. Ackermann et al. (1970). Grün (1982) obtained very high accuracies by using self-calibrating bundle block adjustment based on imagery captured by analog aerial cameras.

Direct georeferencing, often also named direct sensor orientation, is an appropriate alternative if the mobile platform is equipped with a high-quality GNSS/INS system. In order to provide camera poses, a Kalman filter processes the complementary GNSS and inertial observations. While IMUs are prone to drift effects, GNSS updates that are available at a lower frequency can stabilize the combined solution. However, the resulting quality is also dependent on the system calibration, initialization and trajectory characteristics. The concept of direct georeferencing was introduced by Schwarz et al. (1993) and extensively investigated by Skaloud (1999) and Cramer (2001). Since no GCPs are utilized, possible systematic GNSS errors, e.g. due to cycle slips, lead to large mapping deviations in object space. Hence, Cramer (2001) and Haala (2005) proposed to correct systematic offsets as well as time-dependent drifts by incorporating additional image observations.

Combining direct and indirect georeferencing leads to **integrated georeferencing**. In addition to position and attitude information provided by a GNSS/INS system, it employs image observations of tie points and optionally ground control points. Heipke et al. (2002) present a comparison of several approaches in terms of direct and integrated georeferencing. Incorporating only one GCP into integrated georeferencing resulted in object space accuracies similar to conventional aerial triangulation using numerous GCPs. Integrated georeferencing is particularly beneficial in challenging areas with moderate GNSS coverage or in case of low-quality GNSS/INS modules. This applies to UAV platforms with limited payload capacity and thus low-cost navigation sensors. Eugster (2011) presents approaches that enable real-time georegistration of video streams and single images captured by UAVs based on coarse initial position and attitude information in combination with existing digital 3D city models. Rehak (2017) shows that integrated georeferencing allows for UAV mapping with accuracies at the centimeter level even by not utilizing any GCPs. Instead of employing position and attitude weighted observations from a separate inertial/GNSS fusion step, Cucci et al. (2017) directly include raw inertial observations into the bundle adjustment of a dynamic network. Alternatively, initial camera poses can also be derived from a SLAM procedure.

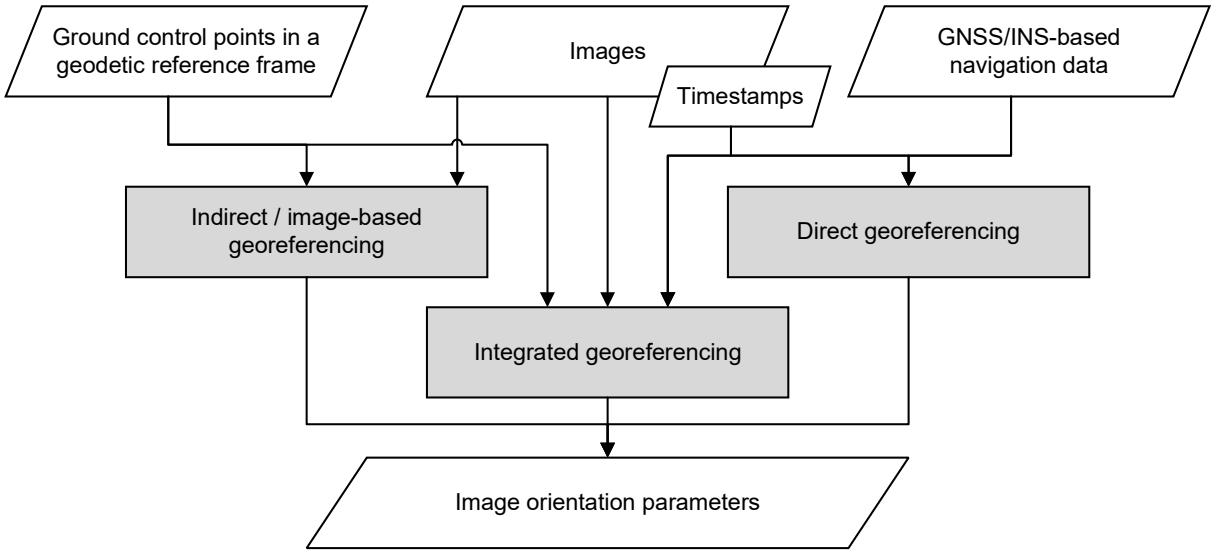

Figure 2.1: Comparison of the three different photogrammetric image orientation approaches (adapted from Eugster (2011)).

2.1.2 Structure from Motion (SfM)

Estimation of camera poses and reconstruction of sparse 3D scene structures is tackled in the computer vision community by structure from motion. While relying on projective geometry, no initial values are required. We distinguish two main concepts: incremental and global approaches. Both procedures expect image correspondences, which are determined by feature extraction and matching combined with geometric verification (see figure 2.2). In case of incremental SfM (Wu, 2013), a carefully selected two-view

reconstruction seeds the model. Subsequently, the procedure incrementally registers new images, triangulates scene points, refines the reconstruction using bundle adjustment and filters outliers (Schönberger and Frahm, 2016). In order to obtain accurate results, this sequential process implies repeated bundle adjustment that is time-consuming. In contrast, global SfM approaches are inherently parallelizable and only require a single bundle adjustment (Sweeney et al., 2015; Reich et al., 2017). All relative poses among image pairs or triplets are considered to simultaneously estimate all camera poses in a single step. Camera orientations are estimated by rotation averaging and camera positions by translation averaging. Tron et al. (2016) give a survey of promising solutions for robust rotation optimization and they benchmark the robustness against the presence of outliers. Cui and Tan (2015) improved the harder problem of translation averaging by using depth maps.

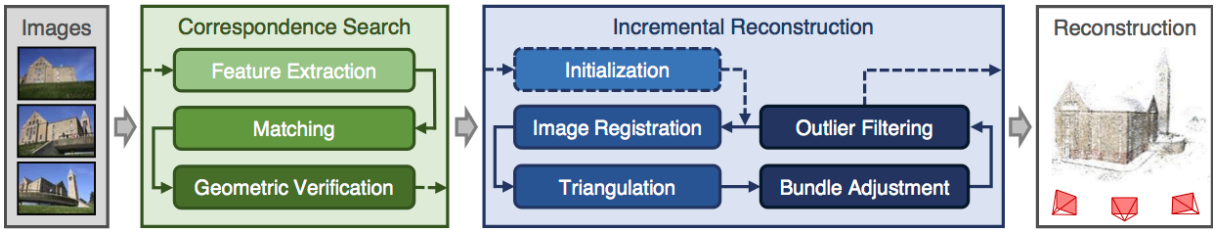

Figure 2.2: Standard processing pipeline of COLMAP for image orientation that is an incremental structure-from-motion approach (Schönberger and Frahm, 2016).

Conventional incremental and global SfM approaches are not ideally suited to reconstruct large-scale image collections. Aiming at improved scalability and efficiency, large SfM problems are typically divided into multiple better conditioned sub-problems that are optimized independently. The resulting sub-models are frequently merged by finding common 3D points across the models and by robustly estimating a similarity transformation using a RANSAC method (Gherardi et al., 2010; Havlena et al., 2010; Parys and Schilling, 2012). Klingner et al. (2013) solve concurrent bundle adjustment problems with fewer than 1500 images and then match the 3D models to each other in order to obtain one consistent reconstruction. Toldo et al. (2015) perform hierarchical SfM and describe the entire SfM process as a binary tree constructed by image clustering. Each leaf corresponds to a single image, while internal nodes represent partial models obtained by merging two sub-nodes. Computation proceeds from bottom to top, starting from several seed couples and eventually reaching the root node. This scheme is able to cut the computational complexity by one order of magnitude and it is less sensitive to both initialization and drift. Partitioning into smaller instances and hierarchical combination makes the problem inherently parallelizable (Toldo et al., 2015). Shah et al. (2015) give a good overview of approaches for sub-tasks of large-scale SfM. They propose a multistage approach that initially generates a coarse 3D model based on a subset of images. This initial solution allows for efficient orientation of the remaining images as well as feature matching and triangulation. Zhu et al. (2018) present a distributed framework that significantly enhances the efficiency and robustness of large-scale motion averaging. They first divide all images into multiple partitions that preserve strong data association for well-posed and parallel local motion averaging. Then, they solve global motion averaging that determines image poses at partition boundaries and a similarity transformation per partition to register all images in a single coordinate frame. Finally, local and global motion averaging are iterated until convergence. In order to improve computational efficiency, Cui et al. (2017) developed a hybrid SfM approach. An adaptive rotation averaging method first estimates camera rotations in a global manner. Based on these, camera centers are computed incrementally. Global rotation averaging decreases the risk of scene drift, and incremental estimation of image positions shows improved robustness in case of noisy data. Cui et al. (2019) present an efficient and robust incremental reconstruction system. While careful track selection significantly reduces the number of feature tracks used in bundle adjustment, an improved rotation averaging method enables to reconstruct both general and ambiguous image datasets. In contrast, Zhao et al. (2018) use a hierarchical approach to solve SfM problems by decoupling the linear and non-linear components. The algorithm begins with small local reconstructions based on non-linear bundle adjustment. These are then merged in a hierarchical manner using a strategy that requires to solve a linear least squares optimization

problem followed by a non-linear transform. Linearity has the advantageous properties that it does not rely on initial estimates, there is no need for iterations and the solution will not end in a local minimum.

2.1.3 Simultaneous Localization and Mapping (SLAM)

SLAM procedures aim at simultaneous sensor pose estimation and map generation of the environment. A broad overview of the current state of SLAM is given by Cadena et al. (2016). SLAM methods can basically be subdivided into visual SLAM, visual odometry and lidar SLAM. In case of lidar SLAM, captured point clouds are continuously aligned and laser scanner poses determined (Hess et al., 2016). However, cameras are often preferred due to their low costs and high frame rates. Visual odometry allows for relative pose estimation of consecutive views by tracking image information in real-time (Scaramuzza and Fraundorfer, 2011; Fraundorfer and Scaramuzza, 2012). Additionally incorporating IMU observations leads to visual-inertial odometry algorithms (Lynen et al., 2015; Forster et al., 2016). Visual odometry focus on local consistency of the trajectory and a local map is used to obtain a more accurate estimate of the local trajectory. In contrast, visual SLAM is concerned with global map consistency. Hence, visual odometry can be used as a building block for a complete SLAM algorithm to recover the incremental motion of a camera. Further components include loop closure detection and possibly a global optimization step to obtain a metrically consistent map. Loop closure is the process of observing the same scene by non-adjacent frames and adding a constraint between them, which considerably reduces the accumulated drift in the pose estimation (Yousif et al., 2015).

SLAM approaches mainly focus on robustness and speed, and they basically consist of a frontend and a backend. The frontend is responsible for feature extraction and matching, outlier removal as well as loop closure detection. The backend performs pose and structure optimization by tools such as g2o (Kümmerle et al., 2011) or GTSAM (Dellaert, 2012) that are based on factor graphs. A factor graph is a bipartite graph consisting of factors connected to variables. These variables represent the unknown random variables in the estimation problem, while the factors represent probabilistic information on those variables, derived from measurements or prior knowledge. SLAM approaches have been categorized into filtering approaches and smoothing approaches (Yousif et al., 2015). Filtering methods solve online SLAM problems by incorporating sensor measurements as they become available. In contrast, smoothing procedures address the full SLAM problem by typically using a least squares error minimization technique. Recent results in monocular visual-inertial navigation have shown that full smoothing approaches based on non-linear optimization outperform filtering methods in terms of accuracy, since they are able to relinearize past states (Forster et al., 2016). However, this improvement comes at the cost of increased computational complexity.

2.1.4 Comparison of Pose Estimation Approaches

Depending on the originating community, image orientation approaches feature different characteristics (see table 2.1). Integrated georeferencing uses collinearity equations to employ tie point and GCP as well as initial EOP observations. Linearized observation equations demand initial values, which are not required by SfM or SLAM. SfM relies on projective geometry and thus utilizes homogeneous coordinates. It incorporates solely image feature correspondences that lead to arbitrary scaled reconstructions. In order to obtain a spatially correct solution, a subsequent 3D similarity transformation is needed. Visual SLAM basically performs feature tracking in sequential images. If a stereo camera configuration or additional IMU observations are available, metric relative camera poses can be computed. In contrast, integrated georeferencing provides camera poses in a global geodetic reference frame. Real-time performance is of utmost interest for SLAM methods, thus cameras with low geometric resolutions but with high frame rates are preferred. Airborne mapping featuring high geometric image resolution, regular flight patterns and high image overlap is a typical application for indirect georeferencing, often also referred to as aerial triangulation or bundle block adjustment. All approaches contain a component that performs global refinement using bundle adjustment, which is described in section 2.1.5.

	Indirect / integrated georeferencing (IG)	Structure from motion (SfM)	Visual SLAM
Originating community	photogrammetry	computer vision	robotics
Approaches	collinearity equations	projective geometry	filtering or smoothing
Observations	tie points, GCPs, initial EOPs for IG	tie points	tie points
Initial values	needed	not required	not required
Coordinate system	global geodetic	local	local
Processing	offline	offline	real-time
Geometric resolution	high	medium	low
Image configuration	regular pattern	arbitrary	sequence
Overlap	high	medium	high
Typical applications	airborne mapping	close-range	navigation

Table 2.1: Comparison of characteristics for three different image orientation approaches.

2.1.5 Bundle Adjustment

Based on a set of measured image feature locations and correspondences, bundle adjustment (Triggs et al., 2000) aims to refine both 3D tie point positions and camera parameters by minimizing reprojection errors. This optimization is usually formulated as a non-linear least squares problem, where the error is the squared L2 norm of the difference between the observed feature location and the projection of the corresponding 3D tie point on the image plane of the camera (Agarwal et al., 2010). For the following paragraph that shows how to solve non-linear least squares problems by the popular Levenberg-Marquardt algorithm, we reference to Agarwal et al. (2010) and Agarwal et al. (2020). Let $x \in \mathbb{R}^n$ be an n-dimensional vector of variables, and $F(x) = [f_1(x), ..., f_m(x)]^\top$ be an m-dimensional function of x. We intend to solve the following optimization problem:

$$\min_x \frac{1}{2}\|F(x)\|^2 \tag{2.1}$$

The Jacobian matrix $J(x)$ of $F(x)$ is an $m \times n$ matrix, where $J_{ij}(x) = \partial_j f_i(x)$ and the gradient vector $g(x) = \nabla \frac{1}{2}\|F(x)\|^2 = J(x)^\top F(x)$. The general strategy for solving non-linear optimization problems is to solve a sequence of approximations to the original problem. At each iteration, the approximation is solved to determine a correction Δx to the vector x. In case of non-linear least squares, an approximation can be defined by using the linearization $F(x+\Delta x) \approx F(x) + J(x)\Delta x$, which leads to the following linear least squares problem:

$$\min_{\Delta x} \frac{1}{2}\|J(x)\Delta x + F(x)\|^2 \tag{2.2}$$

However, iteratively solving a sequence of these problems leads to an algorithm that may not converge. To avoid this, the size of the step Δx needs to be controlled, which can be done by introducing a regularization term:

$$\min_{\Delta x} \frac{1}{2}\|J(x)\Delta x + F(x)\|^2 + \lambda\|D(x)\Delta x\|^2 \tag{2.3}$$

Here, $D(x)$ is a non-negative diagonal matrix and λ is a non-negative parameter that controls the strength of regularization. This damping factor λ is adapted until a decrease of the objective function is reached, which enables Levenberg-Marquardt to gradually switch between gradient descent and the Gauss-Newton method (Sünderhauf, 2012). Gradient descent continuously moves by small steps into the direction of the negative gradient until convergence. While it guarantees to always decrease the value of the objective function and thus converges to a minimum, its performance is rather slow. On the other hand, Gauss-Newton converges very fast (quadratically) when already close to the solution, but it can

fail to give a valid descending step when the optimizer is still far from the sought minimum (Sünderhauf, 2012).

SfM and SLAM procedures rely on feature extraction and matching, which are prone to outliers. A significant number of these can be eliminated by robust estimation methods such as RANSAC (Fischler and Bolles, 1981) based on the epipolar geometry. However, loss functions are frequently essential in order to reduce the influence of noisy data on the solution of non-linear least squares problems. Irschara (2012) evaluates the performance of different robust loss functions such as Huber and Cauchy for bundle adjustment.

Large-scale datasets lead to huge numbers of variables in the optimization problem, since each camera has 6 DoF and each 3D point position 3 DoF. However, for almost all problems, the number of cameras is much smaller than the number of points. Furthermore, feature correspondences are limited to adjacent images. Hence, exploitation of the natural structure and sparsity of the bundle adjustment problem is crucial for numerically stable and efficient procedures. The sparse bundle adjustment approach of Frahm et al. (2010) utilizes sparse Cholesky decomposition methods in combination with a suitable column reordering scheme. Similarly, Schönberger et al. (2014) employ sparse Cholesky factorization and maximize sparsity structure by column and row reordering, but they additionally use the Schur complement trick (Triggs et al., 2000; Agarwal et al., 2010). Its purpose is to first solve the reduced camera system and then to update the points via back-substitution.

Processing large image blocks can also be facilitated by reducing the number of tie points (Karel et al., 2016) or even not relying on any 3D points, which is structureless bundle adjustment. For such approaches, the optimization cost is not based on reprojection errors but on multiple view relations such as the epipolar or trifocal constraints (Steffen et al., 2012; Rodríguez López, 2013; Cefalu and Fritsch, 2014; Indelman et al., 2015; Zheng and Wu, 2015). The structureless procedure developed by Cefalu et al. (2016) allows for self-calibration of camera parameters and the authors mention the main drawback of structureless bundle adjustment. Constraints describe relations between observations, so that the number of equations easily exceed that of classical bundle adjustment. This is especially the case for highly redundant scenarios where the same object points are observed multiple times.

Discussion

We aim at an accurate and robust image orientation procedure. SLAM methods perform in real-time, but they are not able to meet our high accuracy requirements. SfM approaches show good performance, support arbitrary image collections and they are scalable. However, SfM delivers unknown scene scales and thus non-metric reconstructions. In contrast, integrated georeferencing provides accurate camera poses in a predefined coordinate reference frame, but requires initial values. A combination of these two approaches seems to be promising, i.e. extending an SfM approach with georeferencing capabilities.

2.2 Image Orientation for Street-Level and Indoor Multi-Camera Systems

Multi-camera mobile mapping systems allow for efficient data capture in both outdoor and indoor environments. While poor GNSS coverage in urban canyons leads to reduced image orientation accuracies, there is even no GNSS availability in indoor spaces. Hence, image-based georeferencing approaches are essential. Subsequent steps such as dense image matching demand sub-pixel accurate camera poses. Largely varying camera views pose a challenge for establishing tie point connections. However, an adequate number of such correspondences are crucial in order to obtain satisfying bundle adjustment results. Therefore, exploiting relative orientation constraints among cameras is a key element.

2.2.1 Pose Estimation of Multi-Camera Systems

Several image-based mobile mapping systems feature at least two pinhole stereo cameras pointing forward (Novak, 1991; Schwarz et al., 1993; Burkhard et al., 2012). In order to obtain larger coverages, multiple stereo camera systems (Meilland et al., 2015; Cavegn and Haala, 2016) or hybrid configurations consisting of both stereo and panorama cameras in combination with lidar sensors are employed (Paparoditis et al., 2012). Blaser et al. (2018) assembled a 360° stereo configuration based on forward looking pinhole cameras and two tilted panorama cameras featuring fisheye optics (Abraham and Förstner, 2005; Schneider et al., 2009). In contrast, Van Den Heuvel et al. (2006) present a mobile mapping system with a single 360° camera configuration. To perform 3D mapping, they rely on panorama imagery captured at two different vehicle positions, i.e. virtual stereo bases. While efficient, such an approach cannot provide the high relative accuracies that are achieved by rigid and precalibrated physical stereo bases.

Schneider et al. (2012) developed a bundle adjustment approach for omnidirectional and multi-view cameras. Such configurations allow rays perpendicular to the viewing direction, which cannot be transformed into image coordinates for the collinearity equations. Furthermore, the collinearity equations are not suited for far points, since small angles between rays lead to numerical instabilities or singularities. Hence, they use a minimal representation of homogeneous coordinates for image and scene points, enabling scene points at the horizon to significantly stabilize camera rotations. Same as Schneider et al. (2012), Kneip et al. (2013) use bearing vectors (camera rays) instead of image point coordinates as observations. They propose an efficient non-iterative approach for absolute pose estimation that does not require any initial values. For evaluation, the following camera configurations are used: one forward and one backward looking camera, two stereo cameras directed forward, one forward and one sideward pointing camera, forward and backward as well as two sideward looking cameras. Employing camera systems pointing into all four viewing directions led to the best results, which was confirmed by Urban et al. (2017). In contrast to Schneider et al. (2012) and Kneip et al. (2013), they extended the common collinearity equations with a general camera model supporting arbitrary multi-camera systems. Their generic and modular bundle adjustment method does not only allow for pose estimation, but also for simultaneous self-calibration and scene reconstruction. Kersting et al. (2012) modified the conventional collinearity equations to incorporate GNSS/INS-derived positions and orientations as well as relative sensor orientations. Their single-step calibration method for multi-camera mobile mapping systems has the ability to estimate two sets of ROPs, namely the lever arm offsets and the boresight angles relating the cameras and the IMU body frame as well as the ROPs among the cameras.

Havlena et al. (2008) adapted their SfM approach in order to process imagery captured by a stereo rig with two omnidirectional cameras. Employing stereo constraints improved the stability of the reconstruction and helped to keep its overall scale. SfM without constraining the stereo base worked sufficiently when using additional GNSS/INS data, but failed when such data was not available. While many open-source SfM pipelines support both perspective and fisheye camera models, they do not allow for the incorporation of relative orientation constraints (Sweeney et al., 2015; Moulon et al., 2017; Schönberger and Frahm, 2016; Mapillary, 2020). However, OpenMVG (Moulon et al., 2017) and OpenSfM (Mapillary, 2020) incorporate prior camera positions and ground control points. This is also the case for Rumpler et al. (2017) who developed a complete processing pipeline from image capturing to mesh generation. Pix4D[1] and Agisoft[2] meanwhile offer commercial photogrammetric software products enabling multi-camera rig processing including ROP self-calibration. Similarly, the open-source photogrammetric software MicMac (Rupnik et al., 2017) allows constraining relative orientation parameters among cameras as well as incorporating camera positions and ground control points.

For airborne mapping scenarios, oblique images from the individual camera heads are usually treated independently in aerial triangulation (Cavegn et al., 2014; Rupnik et al., 2015; Karel et al., 2016). However, Sun et al. (2016) parametrize oblique camera poses with nadir camera poses as well as constant relative poses between oblique and nadir cameras. This leads to a decrease of the number of unknown parameters and the dimension of the normal equations, which dramatically reduces the computational complexity and memory cost. Relative observations that relate the position and attitude parameters of two consecutive epochs are exploited by Rehak (2017) as well as by Schönberger et al. (2014) in the case

[1] https://www.pix4d.com
[2] https://www.agisoft.com

of UAV mapping. Klingner et al. (2013) optimize trajectories from Google Street View (Anguelov et al., 2010) cars by leveraging precise relative poses from IMU and by considering a generalized camera model that supports rolling shutter cameras. Liu et al. (2018) developed a visual odometry algorithm for a street-based multi-camera system. Their method is particularly more robust in challenging environments, in which a single stereo configuration easily fails due to the lack of features. In order to provide accurate and reliable vehicle localization in real-time, Geppert et al. (2019) fuse absolute pose estimates with relative motion estimates from a multi-camera visual-inertial odometry pipeline. Not limiting to a specific camera configuration, Kuo et al. (2020) present an adaptive SLAM system that works for arbitrary multi-camera setups.

2.2.2 Georeferencing Approaches for Mobile Platforms in Urban Canyons

Street-based mobile mapping systems enable absolute 3D measurement accuracies of a few centimeters in open areas that feature good GNSS conditions (Barber et al., 2008; Haala et al., 2008; Burkhard et al., 2012). However, high-rise buildings forming urban canyons degrade GNSS coverage and there are GNSS multipath effects, which result in poor direct georeferencing accuracies. Hence, integrated georeferencing is essential in challenging built-up urban environments. This even allows for accurate image orientation of data captured at different days and daytimes, which is typical for city-wide mobile mapping.

Ellum and El-Sheimy (2006) proposed to feed coordinate updates (CUPTs) determined by photogrammetric bundle adjustment back into a loosely coupled GNSS Kalman filter. This approach incorporating additional stereovision-based position updates was later on exploited by Eugster et al. (2012). While they demonstrated a consistent improvement of the absolute 3D measurement accuracy from several decimeters to a level of 5-10 cm for land-based mobile mapping, Ellum and El-Sheimy (2006) achieved no improvement in mapping accuracy. Bayoud (2006) developed a SLAM system that does not rely on GNSS observations, but is solely based on inertial observations and tie points from vision sensors. Vision updates for position and orientation are employed as external measurements in an inertial Kalman filter. Forward intersection based on the resulting filtered poses allows for 3D coordinate computation of surrounding features, which can be used as control points for resection in the next epoch. Hassan et al. (2006) perform bundle adjustment incorporating camera positions and orientations provided by a Kalman filter. In areas with poor GNSS coverage, weights of camera positions and orientations are small and hence the solution will only depend on image observations, which results in photogrammetric bridging. Similar approaches were also developed by Forlani et al. (2005) and Silva et al. (2014) in order to bridge street-based mobile mapping stereo image sequences in GNSS denied areas.

Brenner and Hofmann (2012) use 3D landmarks to georeference their lidar mobile mapping system in challenging environments. They automatically extract pole-like structures and match them to accurately positioned reference objects. However, the number of such landmarks is limited in real-world environments, and they struggle with false positive detections. Instead of pole objects, Tournaire et al. (2006) extract and reconstruct crosswalks in order to co-register street-level mobile mapping data and aerial imagery. The motivation is that airborne surveys are much less affected by GNSS degradations experienced by ground-based mobile mapping systems in challenging urban environments. They show the feasibility of obtaining stereo camera poses with sub-decimeter accuracy by relying on an external aerial reference and utilizing ground control objects. Similarly, Javanmardi et al. (2017) use road marking correspondences to correct lidar mobile mapping data by airborne nadir images. Their proposed framework only performs a two-dimensional registration and they report a 2D accuracy of 12 cm. In order to avoid the costly process of GCP determination, also Nebiker et al. (2012) proposed to fuse ground-based imagery from mobile mapping systems with airborne nadir images. First experiments resulted in horizontal accuracies in the order of 5 cm, equivalent to the ground sampling distance of the aerial imagery, and vertical accuracies of approx. 10 cm.

Shan et al. (2014) developed a fully automated pipeline to georeference ground-based multi-view stereo models by oblique airborne images. In order to handle large viewpoint variations, they warp terrestrial images into aerial views using depth map information from the MVS models and corresponding camera poses. Then, the synthesized views are matched with aerial images by SIFT feature correspondences, leading to pixel-level accuracies. Wu et al. (2018) perform rectification of both terrestrial and oblique aerial images based on building façades. Feature matching and subsequent combined bundle adjustment

of terrestrial and UAV images resulted in a 3D RMSE of 83 mm. Also Fanta-Jende et al. (2019) mainly use feature correspondences on façades to automatically co-register panoramic mobile mapping and oblique aerial imagery with pixel-level accuracy. Their earlier approach to improve georeferencing of road-based mobile mapping in GNSS denied urban environments employs airborne nadir images as external reference (Jende et al., 2018). They ortho-project panoramic images to an artificial ground plane in object space so that co-registration can be performed. Due to a flying height of ca. 4500 m, they only show a horizontal trajectory improvement. Furthermore, a GSD of 10 cm allows for sub-pixel and thus accurate feature correspondences of salient road markings, but barely enables absolute sub-decimeter accuracies of the mobile platform. Based on the same aerial nadir images, Hussnain et al. (2018) correct trajectories of a mobile laser scanning platform. Their adjustment technique, which is based on B-splines and uses feature correspondences as well as IMU observations, delivers absolute 3D accuracies of ca. 20 cm. According to Hussnain et al. (2019), matching and triangulating road marking features between aerial images provide reference 3D tie point coordinates. Also determining these 2D features in projected point cloud patches allows for trajectory improvement of the mobile mapping system.

Molina et al. (2016) introduce the concept of mapKITE, which is ideally suited for 3D corridor mapping as it combines road-based mobile data capture with UAV image acquisition. A coded optical target placed on top of a terrestrial lidar platform enables continuous correspondences and thus serves as kinematic GCP for UAV sensor orientation. Employing two GCPs at both ends of a segment that measures 2.3 km as well as one kinematic GCP per image resulted in an absolute 2D accuracy of ca. 5 cm and a 3D accuracy of approx. 10 cm (Molina et al., 2017). This approach can also be used the other way around, i.e. trajectory adjustment of street-based MMS in urban canyons.

Discussion

We address the problem of computing accurate image orientations for multi-camera mobile mapping systems in challenging urban environments. Several recent approaches utilize additional aerial products as external reference. However, limiting factors are a typical GSD of 10 cm for airborne surveys and low-cost sensor orientation modules for UAVs. Our mobile platforms that feature multiple cameras are equipped with navigation systems of high quality. Hence, we prefer to use camera poses provided by direct georeferencing as initial values. In combination with relative pose constraints among cameras, these prior EOPs allow to reduce the costly process of GCP coordinate determination on-site to a minimum. Furthermore, constraining ROPs is a crucial factor for accurate and robust georeferencing of highly redundant multi-view image sequences.

2.3 Dense Reconstruction of 3D Urban Scenes

A typical image-based 3D reconstruction pipeline presumes accurate image orientations to compute depth maps, followed by 3D point cloud and 3D surface generation. We focus on image sequences recorded by street-based mobile mapping systems featuring stereo and multi-view stereo camera configurations. Regarding our overall goal of obtaining geospatial 3D image spaces of high quality, we aim at precise, reliable and dense depth maps. For this purpose, imagery captured at multiple epochs needs to be exploited. However, vehicle motion predominantly in camera viewing direction poses some challenges on the stereo matching process. In addition to 3D computation approaches, image matching strategies for scenarios ranging from airborne nadir and oblique through UAV to close-range and street-level mapping are reviewed within the following sections.

2.3.1 Approaches for Image-Based 3D Reconstruction

Dense 3D reconstruction basically comprises three main steps, namely depth map generation, point cloud computation and 3D surface generation. According to Nebiker et al. (2010), each of these three processes

stands for a different 3D city modeling paradigm, i.e. image-based modeling, rich point cloud based modeling, geometric 3D modeling. In case of image-based city modeling, the 3D urban environment is represented by georeferenced 2D images or videos. Such imagery can efficiently be captured by ground-based mobile mapping systems. Nebiker et al. (2015) extend this model that only relies on radiometry by depth information, leading to geospatial 3D image spaces. Hence, such RGB-D imagery allows for a high-fidelity metric photographic representation of the environment. Dense depth maps ideally provide a depth value for each pixel of the corresponding RGB image. Since they are based on the same source, spatial and temporal coherence of the RGB and the depth information can be ensured. This is typically not possible for dense 3D point clouds, since geometry is acquired by laser scanners and radiometry by cameras with different viewing geometries and at different epochs.

Geometric 3D models can basically be subdivided into surface-based and volumetric-based scene representations. Volumetric models may be generated by varying primitives, but they are frequently defined by an uniform arrangement of cubic volumes, called 3D voxels. Surfaces are usually defined by 3D meshes, consisting of planar faces that share the same edges. Triangular and topologically consistent surfaces are well supported by many photogrammetric pipelines, but in particular by visualization tools. Berger et al. (2014) compare numerous surface reconstruction methods that are based on point clouds, while Musialski et al. (2013) provide a comprehensive overview of urban reconstruction.

Computation of 3D scenes fully depends on accurate image orientations (see figure 2.3). These might be provided by direct georeferencing, but challenging environments require integrated georeferencing. Since it incorporates image observations, a sparse scene reconstruction based on distinctive features is additionally generated. Subsequent stereo matching or plane-sweeping stereo allow for depth map computation, while some multi-view matching approaches directly compute 3D point clouds.

Stereo matching includes a rectification and a dense image matching (DIM) process. Image rectification warps a pair of images so that epipolar lines across the two views are parallel. Hence, corresponding points feature identical rows, facilitating DIM methods to only search along the horizontal direction. Dense stereo image matching is often defined as a global optimization problem. It comprises cost calculation of pixelwise matching, followed by aggregation of matching costs, disparity map computation and refinement (Scharstein and Szeliski, 2002; Szeliski, 2011). The cost function usually contains two terms, i.e. a data term and a smoothness term. The data term measures how well the resulting depth map agrees with the input images in terms of radiometry. The smoothness term incorporates the assumption that the scene is piecewise smooth and penalizes assigning different depth values to neighboring pixels. In particular the semi-global matching (SGM) approach introduced by Hirschmüller (2008) and its multiple variants work efficiently, since matching cost aggregation is only performed along 8 or 16 path directions. Nevertheless, a depth value for every pixel covered by a stereo image pair can be computed, and accuracies are at a similar level as the resolution of the imagery. Several pipelines do not only perform two-view stereo matching, but multi-view stereo. In case of Rothermel et al. (2012), a reference image can feature several match images, thus each baseline provides another disparity map for the same view. Hence, disparity map fusion is required, leading to consistent non-redundant 3D geoinformation.

Plane-sweeping stereo was originally proposed by Collins (1996) and successfully applied by Pollefeys et al. (2008) and Gallup (2011) for 3D urban scene reconstruction from street-level imagery. This method tests a set of plane hypotheses and records for each pixel in a reference view the best plane using some dissimilarity measure. It works with an arbitrary number of cameras. Plane-sweeping stereo assumes multiple planes for the depth tests, a reference image, and several match images captured at different camera positions. Images need previously to be corrected for radial distortion, but no image rectification process is required. In order to reduce computation time as well as to improve depth map quality in ambiguous parts of urban scenes, Pollefeys et al. (2008) and Gallup (2011) incorporate plane priors obtained from the reconstructed sparse points. Additional GPU parallel processing leads to real-time computation, which allows to cope with large-scale urban environments.

Multi-view matching approaches do not rely on image spaces, but they directly target the object space. Hence, Bethmann and Luhmann (2017) transferred matching cost calculation and pathwise cost aggregation of the conventional SGM method into object space. Cost calculation is formulated within a dense voxel raster using the gray values of all images instead of pairwise cost calculation in image space. Thus, object-based SGM is not limited to two-view stereo matching, but allows for real multi-image

correlation. Moreover, semi-global pathwise optimization in object space leads to index maps instead of disparity maps. Since index maps directly indicate the 3D positions of the best matches, no fusion procedure is needed. While the benefits of the original SGM method such as robustness in weakly textured areas and good results at sharp object boundaries are maintained, object-based semi-global multi-image matching does not require a previous image rectification process (Bethmann and Luhmann, 2017).

We aim at obtaining precise, reliable and dense depth maps. While depth information is often directly computed, alternative 3D scene representations such as 3D point clouds or 3D meshes need to be reprojected onto the viewing geometry of respective base images (Schär et al., 2018). If not generated based on the same data source, remaining system calibration inaccuracies and moving objects lead to inconsistencies. Nevertheless, such 3D models may feature high-quality geometries, as they can exploit the full potential of the vast number of 3D manipulation and fusion algorithms.

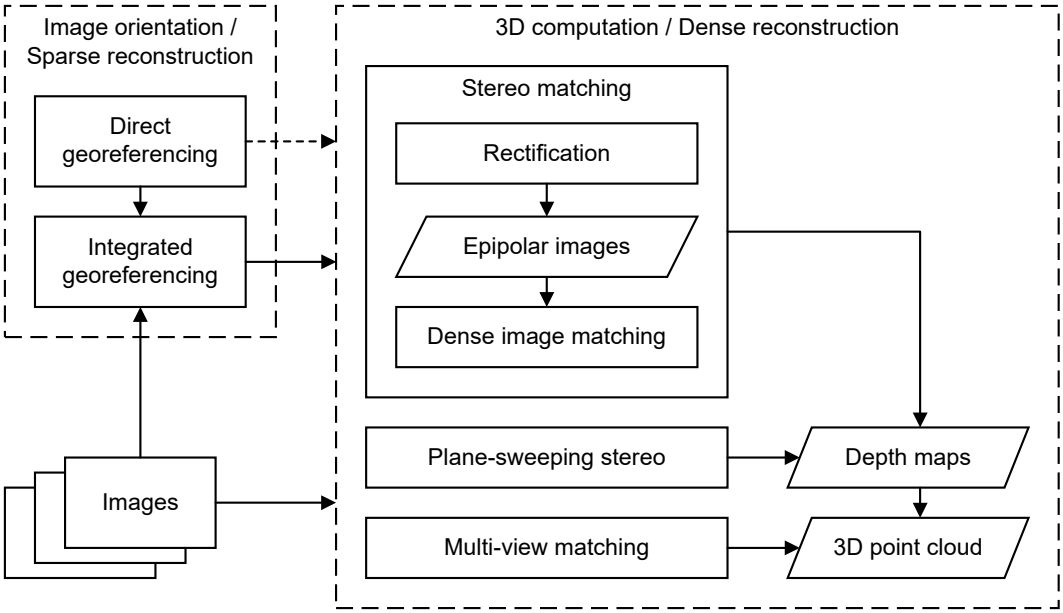

Figure 2.3: Overview of the main processing components for generating 3D geometry solely based on imagery.

Benchmarks have proven to be especially valuable in order to compare varying approaches that address a specified task. Within this section, ongoing benchmarks that do not target a particular application are presented, while benchmarks from the photogrammetric community are described in section 2.3.2. Scharstein and Szeliski (2002) established the first benchmark for two-view stereo correspondence algorithms approx. two decades ago. It initially featured 38 datasets captured by low-resolution cameras. Scharstein et al. (2014) added 33 stereo datasets with a resolution of 6 MP and provided highly accurate ground truth disparities by using a structured light system. However, image acquisition in a controlled indoor environment hardly allows varying conditions. Similarly, Seitz et al. (2006) collected data in a laboratory, but they compare and evaluate multi-view stereo reconstruction algorithms. Only two small objects were mapped by hundreds of images using a low-resolution camera mounted on a precisely controlled robotic arm, while reference data was acquired with a laser stripe scanner. In order to account for real-world conditions, Strecha et al. (2008) introduced a multi-view stereo benchmark featuring outdoor building scenes. It comprises well-textured objects captured by a 6 MP camera, as well as by a laser scanner to obtain dense 3D ground truth data. In contrast, Schöps et al. (2017) and Knapitsch et al. (2017) provide a variety of challenging outdoor as well as complex indoor scenes. The benchmark of Schöps et al. (2017) aims at evaluating both two-view stereo and multi-view stereo algorithms. They captured images using a high-resolution camera (24 MP), and they also employed a synchronized multi-camera rig for collecting low-resolution imagery (0.4 MP). Same as Knapitsch et al. (2017), ground truth scene geometry was acquired using a high-precision laser scanner. However, Knapitsch et al. (2017) focus on high-resolution video sequences as input, encouraging the development of new approaches that leverage

the provided high temporal resolution for increased reconstruction fidelity. Moreover, they allow for the evaluation of complete reconstruction pipelines, i.e. image orientation and dense 3D reconstruction.

2.3.2 Image Matching Strategies for Aerial and Terrestrial Scenarios

Aerial photogrammetry is an established 3D data capturing method for large areas (see table 2.2). In case of airborne nadir scenarios, one large-format aerial camera is directed perpendicular to the flying direction. Large distances to respective mapping areas lead to rather small depth of field variations. Regular flight patterns facilitate both two-view and multi-view stereo matching for the computation of digital surface models (DSMs). Dense pixelwise matching allows for the generation of such results at a resolution that corresponds to the ground sampling distance (GSD) of the original images. However, only one height value per raster cell is usually computed, which leads to 2.5D scene representations.

Haala (2014) presents results of the benchmark on high density image matching for DSM computation. It features two nadir datasets with different land use and block geometry. One dataset includes three strips with 12 images each, captured over a semi-rural area with undulating terrain. With a GSD of 20 cm and a moderate image overlap of ca. 60% in both directions, this dataset is representative for statewide data collection. The other dataset incorporates a densely built-up urban area mapped by three strips with five images each. It has a smaller GSD of 10 cm as well as a higher overlap of 80% in both directions. These two datasets served for the evaluation of eleven dense image matching (DIM) solutions in terms of accuracy and reliability. A common reference surface was obtained by computing a median DSM based on all DIM results. Although independent ground truth is preferred, median DSM surfaces are appropriate to illustrate DIM differences, showing regions that are potentially challenging for DIM. For such areas, elevation profiles from the available DSMs were additionally investigated. There are increased DSM deviations for fine object structures at a size similar to the resolution of the available imagery. A suitable image overlap allows to eliminate erroneous matches and supports the generation of DSMs at vertical accuracies close to the sub-pixel level (Haala, 2014).

In addition to one nadir looking camera head, airborne oblique scenarios usually feature four tilted camera heads pointing in the cardinal directions. These oblique views enable that even building façades and other vertical objects as well as building footprints are represented in the imagery. In contrast to nadir mapping, oblique images allow for true 3D modeling. However, applying DIM algorithms to oblique imagery introduces some major challenges. These include great illumination changes, large perspective deformations and multiple occlusions due to high buildings and vegetation. A higher depth of field leads to larger image scale variations and thus varying GSDs within the same images. Gerke (2009) successfully applied the sophisticated SGM technique to a set of oblique airborne images. He compared the resulting disparity maps with a reference map derived from airborne lidar, but he also computed 3D point cloud differences. Around 60 to 70 percent of all matches were within a range of up to three pixels.

Cavegn et al. (2014) introduced a benchmark dataset aiming at the evaluation of high density image matching based on oblique airborne imagery. It was captured over an urban area by a medium-format camera featuring a nadir and four oblique camera heads tilted by 35°. Nadir images have a GSD of 6 cm, while there is a GSD of 6-13 cm for all four oblique views. The provided oblique aerial image block consists of three strips with nine images each for all five views, resulting in a total of 135 images. The approximate image overlap in nadir view is 70% and 50% for along and across track, respectively. Generated 3D DIM point clouds are compared to reference data collected by terrestrial laser scanning (TLS). Several building façade patches serve for DIM quality evaluation regarding accuracy, reliability and density. Point cloud density is specified as the number of points per square meter. However, higher density values do not necessarily mean better quality, since point cloud filtering may lead to lower density values but better geometry results. Flatness errors indicate the noise level of the extracted 3D geometry and they are calculated based on all point cloud deviations to a best fitting plane. Deviations between DIM results and reference data are evaluated by RMSE, mean values and grid visualizations. Furthermore, profiles reveal the matching resolution, potential systematic errors and accuracies. In order to validate the proposed evaluation procedure, Cavegn et al. (2014) include results from SURE (Rothermel et al., 2012). They overcome the significant increase in disparity search space as well as the resulting higher processing times and memory requirements by employing a modified SGM method called tSGM. It determines the search space for every pixel individually using a pyramid based multi-resolution approach. The evaluation

of a tower building featuring planar façades with several windows and a school building whose façades show distinctive structures resulted in DIM accuracies of ca. 1-2 pixel.

Further oblique benchmarks were introduced by Nex et al. (2015) and Özdemir et al. (2019). In addition to DIM evaluation, they incorporate tasks such as image orientation. Nex et al. (2015) provide 85 oblique and nadir airborne images featuring a GSD of approx. 10 cm and an image overlap of ca. 80% in both directions. Evaluation is based on TLS as reference data and performed according to the procedure developed by Cavegn et al. (2014). Özdemir et al. (2019) captured oblique data of a 3D artifact representing a typical urban scenario by means of a mock 3D city model (ca. 80 cm × 80 cm) in a controlled laboratory environment. They simulated a classical airborne flight with a nadir image overlap of 80% and 65% for along and across-track direction, respectively. Each camera station features six images, i.e. two nadir images and four oblique views tilted by 45°. Upscaling to real-world dimensions would lead to a high GSD of ca. 1 cm. A multi-stripe triangulation-based laser scanner delivered 3D reference point clouds of two buildings. DIM results are evaluated in terms of accuracy and completeness based on patches showing different texture and geometric complexity.

Unmanned aerial vehicles (UAVs) are suitable for 3D mapping of small to medium-scale environments. An extensive overview is given by Colomina and Molina (2014). Compared to airplanes, they are very agile but they have a limited payload capacity, forcing to carry low-weight and thus often low-cost imaging sensors. UAVs allow for data capture of all the three scenarios listed in table 2.2, i.e. area-based as well as object-based and linear mapping. In case of area-based mapping, conventional flight patterns from airborne photogrammetry can be used. In contrast, corridor mapping, such as performed by Molina et al. (2017), does frequently incorporate only one flight strip and not multiple. This requires other image orientation procedures to stabilize imagery in cross direction, and solely enables DIM of images acquired consecutively in flight direction. Object-based mapping leads to varying camera viewing directions and less uniform image collections, which is challenging for DIM. Schär et al. (2018) investigate DIM results based on four datasets, comprising both area-based mapping by a fixed-wing UAV and object-based data capture by multicopter platforms. They incorporate the three DIM solutions SURE (Rothermel et al., 2012), COLMAP (Schönberger et al., 2016) and patch-based multi-view stereo (PMVS) (Furukawa and Ponce, 2010). While SURE is even able to sufficiently reconstruct homogeneous surfaces and COLMAP delivers results with a small noise level, PMVS generates incomplete point clouds. Nex et al. (2015) provide a combined UAV and terrestrial dataset comprising one building for DIM benchmark purposes. The available 228 images have an average GSD of 5 mm and TLS data serves as reference.

Close range photogrammetry is ideally suited for 3D reconstruction of small to medium-sized objects. Similar to airborne nadir and UAV mapping, close-range scenarios typically feature imagery captured by only one camera. However, applying dense multi-view image matching to nonuniform image collections and convergent views poses some challenges. Wenzel (2016) presents a DIM approach and strategies that are able to handle varying image configurations and resolutions. Ahmadabadian et al. (2013) demonstrate the potential of several DIM algorithms to generate accurate and dense point clouds using stereo imagery of four objects. Remondino et al. (2014) perform a comparison of four DIM solutions using three aerial and five close-range terrestrial datasets, featuring variations in image scale, image resolution, image number, camera network, baseline length, object texture and size. Generated dense point clouds are evaluated by a flatness error measure, profiles and Euclidean distances to a TLS mesh.

Multi-camera mobile mapping systems enable efficient street-level image data acquisition. To this end, single stereo, multi-stereo or panoramic camera configurations are often used. In particular cameras looking forward and backward record large scene depth variations, leading to significant image scale changes within the same imagery. Incorporation of images captured at multiple consecutive epochs further aggravates the 3D reconstruction process. Predominant motion in viewing direction between neighboring images results in epipoles located inside or close to the stereo partner. Since conventional rectification approaches (Loop and Zhang, 1999; Fusiello et al., 2000) fail for such scenarios, sequential matching demands advanced rectification procedures. Pollefeys et al. (1999) and Oram (2001) proposed methods that can deal with arbitrary stereo configurations and multiple epochs. Nonetheless, platform movement mainly in camera viewing direction leads to very small baselines, which do not allow stereo matching to generate precise depth maps and accurate 3D point clouds. Geiger et al. (2012) introduced a benchmark featuring stereo image sequences of numerous street sections captured by a mobile mapping

system. Stereo image matching is one of several tasks. A mobile laser scanner delivered accurate ground truth and the provided real-world outdoor scenes reduce the issue of algorithm overfitting. In order to account for moving objects, Menze and Geiger (2015) added another large-scale dataset with ground truth annotations for all static and dynamic scene objects.

	Airborne nadir	Airborne oblique	UAV	Close range	**Street-level**
# cameras	1	usually 5	often 1	1	several
Depth of field	small	medium	medium	medium	**large**
Area-based	x	x	x	(x)	
Object-based			x	x	
Linear			(x)		x
Dimensionality	2.5D	3D	3D	3D	3D

Table 2.2: Characteristics of different image matching scenarios.

Discussion

We aim at accurate 3D urban scene reconstructions based on stereo imagery provided by street-level mobile mapping. In order to exploit the high image redundancy, in-sequence dense image matching needs to be applied. In contrast to airborne nadir scenarios, which feature cameras pointing perpendicular to the flying direction, cameras of mobile mapping systems often face the driving direction. Main platform motion in camera viewing direction aggravates the 3D reconstruction process. Since traditional rectification methods struggle with epipoles located inside or close to the stereo partner, sophisticated approaches such as polar rectification are required.

Summary

Multi-camera mobile mapping systems can provide highly redundant multi-view image sequences. While such imagery covers the surrounding 3D scene very well, establishing feature correspondences among strongly varying views is challenging. Camera pose estimation is even more aggravated in urban environments that do not allow for a sufficient number of GNSS signals. Hence, the image orientation process needs to exploit relative orientation constraints among cameras. Furthermore, incorporation of initial camera poses from direct georeferencing or SLAM as well as ground control points enable accurate georeferencing. Consequently, extending the structure-from-motion pipeline COLMAP with these features is very promising in order to obtain image orientation accuracies at the sub-pixel level. High-quality alignment of multiple image sequences is a prerequisite for successful multi-view stereo matching, leading to accurate 3D scene representations.

Chapter 3

Developed Methods for Integrated Georeferencing

We aim at highly accurate and reliable image orientations in challenging urban environments. For this purpose, redundant multi-view imagery captured by multi-camera systems is exploited. These are either organized as multi-stereo camera rigs or multi-head panorama cameras (see section 4.1.1). Determination of tie point correspondences among images of strongly varying views is challenging and often not feasible. However, constraining relative orientation parameters (ROPs) among cameras mitigates this issue, and just requires the computation of one pose per epoch (see section 3.1). Hence, we not only improved ROP functionality of the powerful SfM pipeline COLMAP, but also implemented georeferencing capabilities (see section 3.2). While prior camera poses from direct georeferencing or SLAM are a prerequisite, incorporation of GCPs allows for a further quality improvement.

We explain our approaches within the next sections using a dataset captured in Basel in July 2014. It features a typical configuration for road mapping in an urban environment. Six pinhole cameras are assembled as three stereo systems, one facing forward, one back-right and one left (see figure 3.1). All sensors are mounted on a rigid frame in order to ensure stability of relative orientations among cameras but also to an IMU. More details concerning the dataset Basel14 can be found in section 4.2.1.

3.1 Relative Orientation Constraints for Image-Based Mobile Mapping

Multi-camera mobile mapping systems allow for efficient data capturing, frequently covering the entire horizontal field of view. Moreover, image acquisition at high frame rates and often in opposite driving directions leads to highly redundant data. Hence, orientation of individual images is a cumbersome approach, which is often not able to provide camera poses for several candidates. However, constraining relative orientation parameters among multiple cameras enables both high accuracy and robustness of image orientation.

Establishing adequate stereo bases is crucial for accurate 3D mapping. As typical in airborne scenarios, images captured by the same camera at two different epochs can build a virtual stereo base. However, such a virtual stereo base contains the inaccuracies of two camera poses derived from georeferencing. Therefore, physical stereo bases are preferably used for applications with shorter distances to objects of interest, which is the case for street-based mobile mapping. As shown by figure 3.1, physical stereo bases require two cameras attached to a rigid frame. Since physical stereo bases can be calibrated with high precision, they allow for 3D accuracies at the millimeter level, while virtual stereo bases rather deliver centimeter accuracies.

The configuration depicted in figure 3.1 features 11 MP cameras for the forward looking stereo system and 2 MP cameras for both the back-right and left pointing stereo systems. The higher geometric resolution in driving direction is due to a larger line of sight compared to sideways as well as the availability

of more objects of interest in the roadway. While this camera configuration that captured the dataset Basel14 enables a stereo coverage in three horizontal directions, other configurations employing multi-head panorama cameras deliver 360° horizontal mappings, and e.g. 270° in the vertical direction. If combined, panoramas are frequently used for navigation and exploration purposes, and measurements for accurate 3D coordinate determination are performed in stereo images. Reasons are small image overlaps for panorama camera heads as well as rather short bases enabling accurate depth estimation only for short distances. In contrast, large image overlaps of individual stereo camera systems allow for many feature matches and thus well connected images due to a similar view and mapped area. However, assigning point connections to other stereovision systems with varying views still poses some challenges.

Figure 3.1: Camera configuration (left) and images captured by all cameras at the same location during a mobile mapping campaign in Basel in July 2014 (right: forward stereo [top], back-right stereo [middle], left stereo [bottom]).

Exploitation of Multi-Camera Rigs

Standard image orientation procedures consider each camera separately. Hence, they would try to estimate six individual poses per epoch in case of our dataset Basel14. While feasible in well-structured environments, these approaches barely deliver accurate and reliable pose estimations for all images in difficult environments with low-texture areas or repetitive patterns. However, since we rely on fixed multi-camera systems, these can technically be defined as multi-camera rigs. Nonetheless, as only feasible for images captured at exactly the same epoch, both a precise synchronization and a stable assembly of all sensors are required. Moreover, the relative orientation parameters among all sensors have to be precalibrated. However, constraining these calibrated ROPs during bundle adjustment means that only reference camera poses need to be estimated, which significantly reduces the degrees of freedom. Since handled as a unit, fewer observations per camera image are necessary, yet preferably distributed all around in order to stabilize the multi-camera rig. The need for fewer feature correspondences is especially beneficial for homogeneous surfaces or moving objects, but also for moderate to non-overlapping FOVs.

ROP Definition and Estimation Procedures

Every multi-camera rig has a reference camera, which is frequently the left camera of the forward pointing stereo system, and varying relative orientation parameters to the other cameras. ROPs between two cameras include three translation and three rotation parameters, that is six in total. In case of our dataset Basel14, either five relations to the reference camera or two relations to the reference camera and three bases can be defined (see figure 3.2). The conventional approach for multi-stereo, which expects position vectors and rotations with regard to a reference camera, was implemented by Schönberger (2020) and is our prime choice for multi-head panorama cameras. Inspired by Kersting et al. (2012), who primarily focused on calibration and not on 3D mapping scenarios, we implemented a sophisticated approach that is best suited for multi-stereo camera configurations. It handles stereo camera systems as individual

units, thus featuring a position vector and a 3D rotation to the left camera and then a base to the corresponding stereo camera partner. Since stereo bases are usually precalibrated with a higher precision than the respective offsets and rotation components related to the reference camera, this approach enables estimating arbitrary single components yet fixing the calibrated stereo values. For many scenarios, self-calibration of rotations among stereovision systems is sufficient.

Our image orientation procedure based on COLMAP facilitates four different options:

- self-calibration of all ROPs
- fixed multi-camera rig (precalibrated ROPs with high precision required)
- fixed bases but estimation of all ROPs among reference camera and stereovision systems (precisely precalibrated stereo bases needed)
- fixed bases but estimation of either offsets or rotations among reference camera and stereovision systems (precisely precalibrated stereo bases needed)

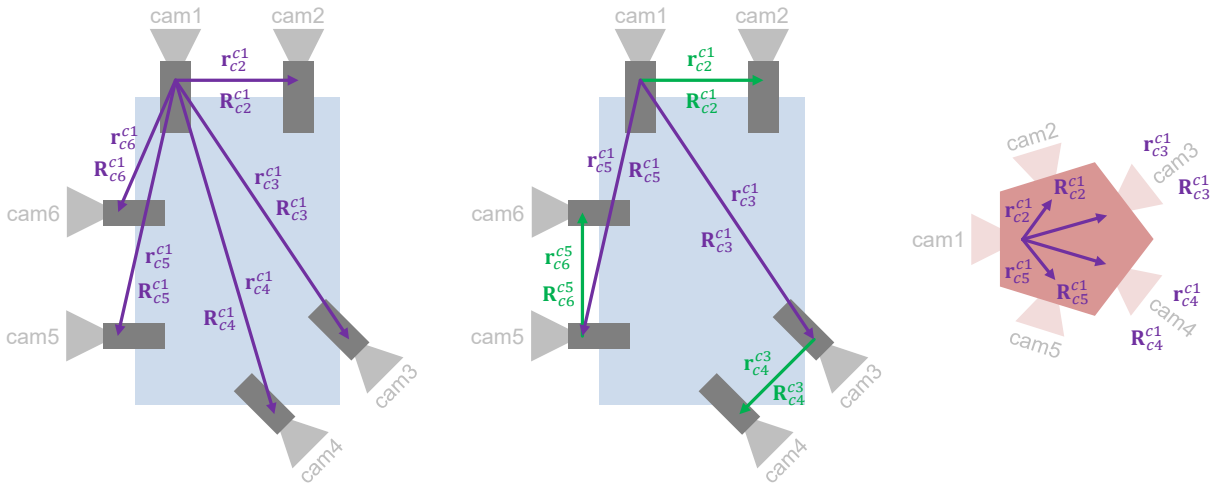

Figure 3.2: Top views of different ROP definition and estimation procedures: conventional approach always referring to a reference camera (left), sophisticated approach that exploits precisely calibrated stereo bases depicted in green (middle), approach for multi-head panorama cameras (right).

3.2 Georeferencing by Prior Camera Poses and Ground Control Points

We aim for photogrammetric reconstructions, thus correct dimensions as well as positioning and attitude information in a predefined geodetic reference frame are of utmost importance. In contrast, proper scaling and georeferencing is often not handled in the computer vision community (see section 2.1.4). Since SfM approaches deliver ambiguous scale factors, either a length reference bar or control points are required in order to obtain metric reconstructions. In case of stereo mapping, the precise length of at least one stereo base is needed.

Georeferencing is frequently achieved by performing a similarity transformation of the SfM model onto GCP coordinates defined within the target datum. Rigid similarity transformations feature 6 degrees of freedom (DoF), namely three translations and three rotations. In case of distorted geodetic reference frames, estimating an additional scale factor is appropriate, which results in 7 DoF and a minimum of three GCPs. However, more accurate results can be obtained by integrated georeferencing. Due to the incorporation of exterior orientation parameters (EOPs) from direct georeferencing or SLAM, potential image block deformations can be reduced or actually avoided. Moreover, exploitation of prior EOPs even

allows for processing image sequences without GCPs. Nonetheless, GCP utilization for highly accurate georeferencing or at least using one check point for validation is recommended.

COLMAP Implementations

The standard COLMAP processing pipeline (see figure 2.2) developed by Schönberger and Frahm (2016) normalizes scenes in order to avoid degenerate visualizations after bundle adjustment and to improve the numerical stability of algorithms. This normalization comprises scene scaling to a predefined extent and scene coordinate reduction to the centroid. Hence, resulting SfM models have unknown scale factors and arbitrary orientations. However, a standard COLMAP post-processing module called *model_aligner* (Schönberger, 2020) enables correct georeferencing by performing a 3D similarity transformation to 3D coordinates of at least three projection centers.

We aimed at a tighter and more accurate georeferencing implementation, which supports the incorporation of EOPs, GCPs and ROPs among cameras. Moreover, ROP definition with both hard and soft constraints was demanded. While the utilization of prior EOPs already delivers scene scale and georeferencing that is typically in the decimeter to meter range, employing GCPs and calibrated ROPs lead to a further refinement. In either case, a weak datum computation is performed, which results in modified 3D coordinates of both projection centers and ground control points.

If GCPs as well as a precise ROP calibration are available, our extended COLMAP procedure carries out integrated georeferencing using a fixed multi-camera rig (see table 3.1, use case I). In case of a moderately precise calibration, bundle adjustment with fixed ROPs among cameras can lead to inconsistencies. Hence, ROP self-calibration is an adequate option (use case II). In order to avoid the time-consuming step of GCP coordinate determination, integrated georeferencing solely based on EOPs and fixed ROPs is the method of choice (use case III). However, ROPs among the reference camera and the other stereo camera systems might be precalibrated with reduced accuracies compared to respective stereo bases. In such a case, the precalibrated length of at least one stereo base needs to be fixed, while the other ROP components can be defined as soft constraints (use case IV). Nonetheless, a precise ROP self-calibration requires accurate initial EOPs as well as redundant image data, preferably captured in opposite driving directions.

Use case	EOPs	GCPs	ROPs	
	Initial		Fixed	Initial
I: IG with GCPs and fixed ROPs	x	x	x	
II: IG with GCPs and ROP self-calibration	x	x		x
III: IG without GCPs but with fixed ROPs	x		x	
IV: IG without GCPs but with ROP self-calibration	x		x	x

Table 3.1: Four use cases with varying exploitation of ground control points (GCPs) and relative orientation parameters (ROPs), but all relying on prior exterior orientation parameters (EOPs). Initial ROPs are precalibrated, and additionally self-calibrated within bundle adjustment.

3.3 Preprocessing Steps for Integrated Georeferencing

Camera poses determined in urban as well as in indoor environments frequently show accuracies in the decimeter range. Reasons are either poor GNSS conditions in urban canyons or camera poses derived from SLAM. In order to improve the accuracy of these prior EOPs, we perform integrated georeferencing, which additionally relies on corrected images, on calibrated ROPs, and optionally on GCPs (see figure 3.3). Hence, high-quality calibration of the complete configuration, an image processing step, and direct georeferencing or SLAM are required. Calibration is an often underestimated yet crucial procedure, playing a key role in ensuring high relative and absolute accuracies. It comprises calibration of interior orientation parameters (IOPs) of each single camera, calibration of offsets and rotations between a reference camera and all the other cameras (see section 3.1), as well as boresight alignment that is the calibration of lever arm and misalignment relating the reference camera and the IMU body frame.

Precisely calibrated interior orientation parameters serve for distortion and principal point corrections of the images.

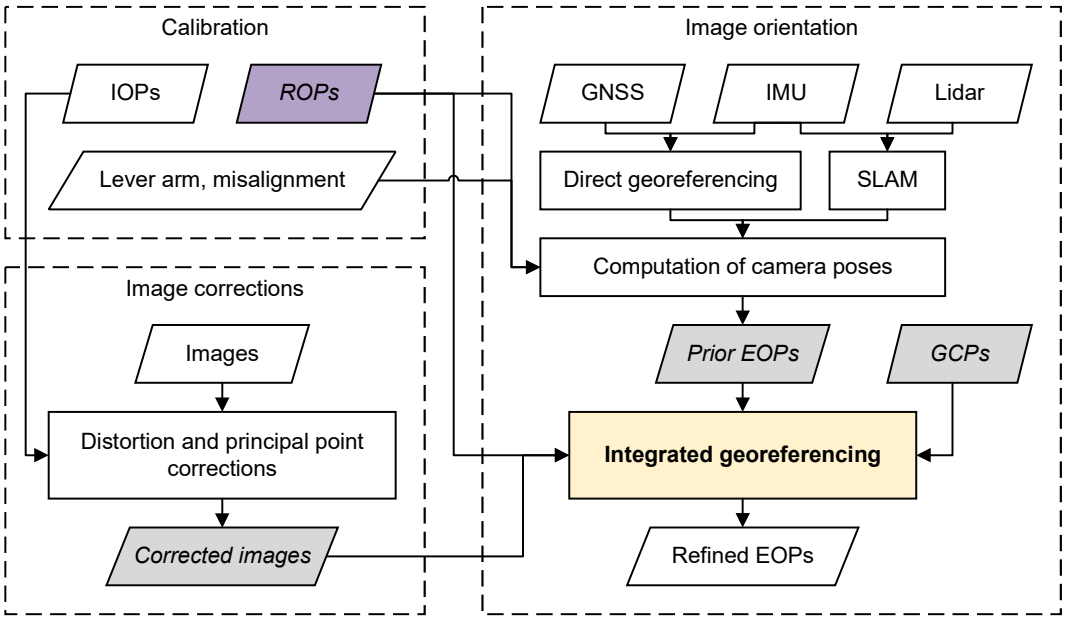

Figure 3.3: Components and processing steps that are needed in order to obtain highly accurate exterior orientation parameters for all images.

3.3.1 Camera and Multi-Sensor System Calibration

Multi-camera mobile mapping systems enable efficient geodata collection for road corridor mapping applications. In order to facilitate the image orientation process, constraining ROPs among individual cameras mounted on a rigid frame is fundamental. Hence, these ROPs but also interior orientation parameters of each camera as well as lever arm and misalignment need to be calibrated precisely, since errors and offsets will be transferred to the full extent to the following 3D mapping steps. While Burkhard et al. (2012) only used one outdoor calibration field for the estimation of all components, we employ two calibration fields for road mapping. An indoor calibration field serves for IOP and ROP estimation, whereas boresight alignment is performed outdoors.

Calibration of Interior and Relative Orientation Parameters

Interior orientation parameters comprise the principal distance c, coordinates of the principal point x_0, y_0 as well as additional parameters for lens distortion, i.e. symmetric radial distortions and asymmetric distortions caused by lens decentering. We usually estimate two radial and two tangential distortion parameters. Our calibration procedure supports both pinhole and fisheye cameras (Blaser et al., 2018). While pinhole cameras follow the perspective projection model, Ladybug5 camera heads feature fisheye lenses that are best fitted by the equidistant projection model (Abraham and Förstner, 2005). As shown in section 3.1, there are ROP estimation approaches for general cases as well as for multi-stereo camera configurations. While the first defines ROPs of all cameras with regard to a reference camera, the latter separates ROPs of stereo bases and ROPs among a reference camera and the respective stereo camera systems. We simultaneously estimate IOPs and ROPs by constrained bundle adjustment using imagery from several epochs captured at different locations in an indoor calibration field (see figure 3.4). Such a field features many signalized coded targets, which are uniformly distributed all around and were determined with sub-millimeter accuracy in 3D. Calibrated ROPs typically show a precision at the millimeter to sub-millimeter level.

Figure 3.4: Indoor calibration field that is utilized for calibrating IOPs and ROPs (Source: iNovitas).

Boresight Alignment

A further step aims at estimating the relative orientation between the reference camera of a multi-camera configuration and the body frame, which is the reference frame of the navigation system (see section 7.1.1). While EOPs of the reference camera are computed by constrained bundle adjustment usually incorporating images captured by the main stereo system pointing forward, body frame EOPs are determined by direct georeferencing. The difference between these two solutions results in a position vector $\mathbf{r}_{c1}^{b}$ and a corresponding rotation $\mathbf{R}_{c1}^{b}$, i.e. lever arm and misalignment (see green components in figure 3.5). Since directly affected by the GNSS/INS solution that often shows accuracies at the centimeter level, boresight alignment contributes a larger part to the total error than IOP and ROP calibration. Therefore, good GNSS coverage at data collection time is essential, which can be ensured in our two outdoor calibration fields. One was established on a basketball court and served for our Basel datasets (see section 4.2.1), while the other features a rather straight road segment. Coordinates of photogrammetric targets and well-defined natural points were determined with a 3D accuracy in the millimeter range.

Please note that we use the following naming convention: A vector $\mathbf{r}$ defines three translations to a target frame (subscript) w.r.t. an original frame (superscript), e.g. $\mathbf{r}_{c1}^{m}$ indicates the position of the camera frame c1 represented in the mapping frame m. A subscript of a rotation matrix defines an original frame, while a superscript stands for a target frame. Therefore, $\mathbf{R}_{c1}^{m}$ indicates a 3D rotation from the camera frame c1 to the mapping frame m.

Camera poses for the dataset Basel14 are computed as follows:

- Pose of reference camera c1 based on body frame pose (see gray components in figure 3.5) as well as on lever arm and misalignment (see green components in figure 3.5):

$$\mathbf{r}_{c1}^{m} = \mathbf{r}_{b}^{m} + \mathbf{R}_{b}^{m}\,\mathbf{r}_{c1}^{b} \tag{3.1}$$

$$\mathbf{R}_{c1}^{m} = \mathbf{R}_{b}^{m}\,\mathbf{R}_{c1}^{b} \tag{3.2}$$

- Poses of cameras c2-c6 based on reference camera (c1) pose and ROPs (see purple components in figure 3.5 as well as in figure 3.2):

$$\mathbf{r}_{c2}^{m} = \mathbf{r}_{c1}^{m} + \mathbf{R}_{c1}^{m}\,\mathbf{r}_{c2}^{c1} \tag{3.3}$$

$$\mathbf{R}_{c2}^{m} = \mathbf{R}_{c1}^{m}\,\mathbf{R}_{c2}^{c1} \tag{3.4}$$

$$\mathbf{r}_{c3}^{m} = \mathbf{r}_{c1}^{m} + \mathbf{R}_{c1}^{m}\,\mathbf{r}_{c3}^{c1} \tag{3.5}$$

$$\mathbf{R}_{c3}^{m} = \mathbf{R}_{c1}^{m}\,\mathbf{R}_{c3}^{c1} \tag{3.6}$$

$$\mathbf{r}_{c4}^{m} = \mathbf{r}_{c1}^{m} + \mathbf{R}_{c1}^{m}\,\mathbf{r}_{c3}^{c1} + \mathbf{R}_{c3}^{m}\,\mathbf{r}_{c4}^{c3} \tag{3.7}$$

$$\mathbf{R}_{c4}^{m} = \mathbf{R}_{c1}^{m}\,\mathbf{R}_{c3}^{c1}\,\mathbf{R}_{c4}^{c3} \tag{3.8}$$

$$\mathbf{r}_{c5}^{m} = \mathbf{r}_{c1}^{m} + \mathbf{R}_{c1}^{m}\,\mathbf{r}_{c5}^{c1} \tag{3.9}$$

$$\mathbf{R}_{c5}^{m} = \mathbf{R}_{c1}^{m}\,\mathbf{R}_{c5}^{c1} \tag{3.10}$$

$$\mathbf{r}_{c6}^{m} = \mathbf{r}_{c1}^{m} + \mathbf{R}_{c1}^{m}\,\mathbf{r}_{c5}^{c1} + \mathbf{R}_{c5}^{m}\,\mathbf{r}_{c6}^{c5} \tag{3.11}$$

$$\mathbf{R}_{c6}^{m} = \mathbf{R}_{c1}^{m}\,\mathbf{R}_{c5}^{c1}\,\mathbf{R}_{c6}^{c5} \tag{3.12}$$

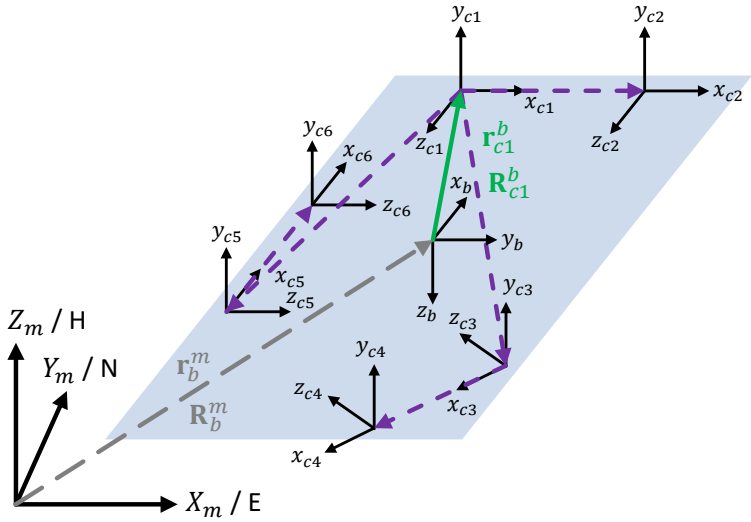

Figure 3.5: Illustration of lever arm and misalignment (green) as well as camera pose computation (purple) including the relevant coordinate systems: mapping frame (m), IMU body frame (b) and coordinate systems of all cameras (c1-c6) for the dataset Basel14.

3.3.2 Direct Georeferencing and SLAM

We utilize cameras as mapping sensors, whereas GNSS and IMU as well as lidar sensors deliver data for initial EOP computation that is required by our integrated georeferencing approach. While outdoor navigation data comprises GNSS and IMU observations suitable for direct georeferencing, lidar and IMU data are processed in a SLAM procedure for indoor environments. In such environments where homogeneous surfaces and repetitive textures are omnipresent, active sensor technologies are preferred over passive and thus the choice for lidar SLAM and not visual SLAM.

Direct Georeferencing

In outdoor environments, we rely on complementary data collected by INS and differential GNSS employing virtual reference stations. While inertial navigation delivers high relative accuracies, GNSS can correct IMU drift effects, enabling accurate global solutions as well (see section 2.1.1). Nonetheless, GNSS signal outages or systematic errors due to multipath effects, incorrectly fixed GNSS carrier phase ambiguities or cycle slips are prevalent in urban environments. We process navigation data in tightly coupled mode using the GNSS and inertial post-processing software Inertial Explorer[1] from NovAtel. Furthermore, we perform processing in multi-pass directions and additionally smooth trajectories. Incorporation of image events with timestamps provides body frame poses in WGS84 and Euler angles as roll γ, pitch θ, heading ψ (see figure 3.6). Subsequently, we transform these global geographical coordinates into a map projected coordinate system as well as ellipsoidal to orthometric or leveled heights. Our application scenarios usually demand a mapping frame of Switzerland, thus a Swisstopo tool named REFRAME[2] can e.g. deliver plane coordinates in the Swiss horizontal reference frames LV03 or LV95, and leveled heights in the Swiss vertical reference frame LN02. Euler angles are converted from the representation roll γ, pitch θ, heading ψ to omega ω, phi φ, kappa κ (see section 7.1.2 for formulas). As explained in section 3.3.1, lever arm (LA) and misalignment (MA) allow for computation of reference camera poses, and utilizing calibrated ROPs among cameras lead to EOPs from direct georeferencing for all images.

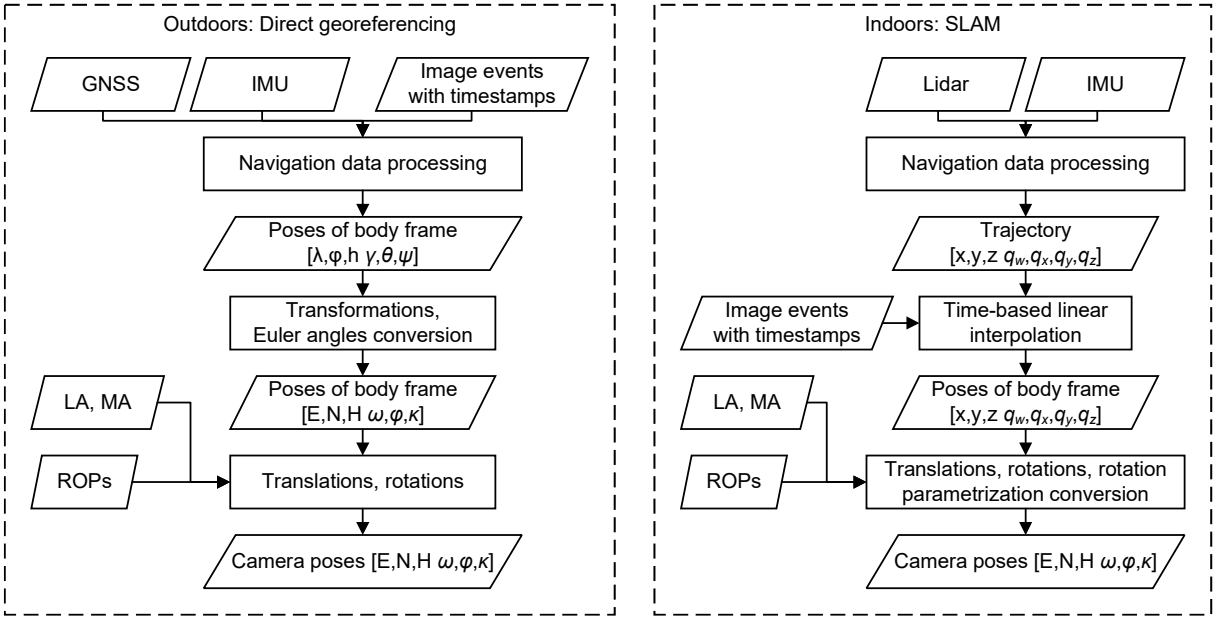

Figure 3.6: Workflows for direct georeferencing (left) and SLAM (right). Resulting EOPs are needed as initial values for our subsequent integrated georeferencing approach.

Simultaneous Localization and Mapping

We perform indoor 3D mapping using a multi-head panorama camera. Nonetheless, trajectories are computed by SLAM (see section 2.1.3) utilizing the Google Cartographer (Hess et al., 2016) that employs lidar and IMU data (see figure 3.6). The starting position of lidar SLAM defines the origin of a local Cartesian coordinate system. Although real-time poses are computed, an optimized sensor trajectory is exported once data collection is completed. In a post-processing step, we conduct time-based interpolation of the camera trigger events between the trajectory events. Linear interpolations for the positions and spherical linear interpolations based on quaternions for the orientations result in body frame poses.

[1] https://novatel.com/products/waypoint-software/inertial-explorer
[2] https://www.swisstopo.admin.ch/en/maps-data-online/calculation-services/reframe.html

Afterwards, we either transform these local poses or control point reference coordinates by a 6 DoF similarity transformation. Moreover, we convert quaternions q_w, q_x, q_y, q_z to the Euler angles representation omega ω, phi φ, kappa κ (see section 7.1.2 for formulas). Employing boresight alignment parameters between the body frame and the reference camera head as well as calibrated ROPs among panorama camera multi-heads result in prior exterior orientation parameters for all images.

3.4 Implementation of our Integrated Georeferencing Approach based on COLMAP

COLMAP is a powerful incremental structure-from-motion tool, which was mainly developed by Johannes Schönberger (Schönberger and Frahm, 2016; Schönberger, 2018). Since established as one of the best performing open-source tools of its kind, COLMAP often serves as a baseline in the computer vision community, representing the pipeline based on hand-crafted features. It supports perspective as well as fisheye camera models. The standard procedure depicted in figure 2.2 first extracts SIFT features, then matches them, followed by geometric verification that leads to a scene graph. Based on a carefully selected two-view reconstruction, the sparse model grows incrementally. The underlying loop consists of image registration, triangulation of scene points, reconstruction refinement by bundle adjustment and outlier filtering. As already mentioned in section 3.2, the resulting SfM models have unknown scene scales and they are not georeferenced. Furthermore, due to strong motion in driving direction and tiny bases, sequences of monocular images barely enable an adequate initialization. While ROPs among multi-cameras can be constrained, i.e. just estimation of all ROPs, precalibrated values cannot be defined and thus none can be fixed. Moreover, there is no GCP support and incorporation of initial EOPs is not possible in the standard COLMAP pipeline. However, Heng et al. (2019) modified COLMAP so that initial pose estimates from a GNSS/INS system can be used for tie point triangulation, which considerably speeds up the process that is especially beneficial for large-scale reconstructions.

We extended COLMAP for the purpose of integrated georeferencing (see section 2.1.1 and figure 3.7). When compared to the standard COLMAP pipeline (see figure 2.2), the main difference is a global and not an incremental reconstruction process (see section 2.1.2). Since we rely on initial EOPs from direct georeferencing or SLAM (see section 3.3.2), we can directly triangulate tie points for all images. Prior EOPs further enable spatial feature matching and they serve for bundle adjustment (see section 2.1.5), where GCPs and ROPs are exploited as well. The standard procedure performs both local and global bundle adjustment. The first is conducted for each new image by employing a predefined number of images, while the latter incorporates all images and is carried out as soon as either the number of registered images or the number of generated 3D tie points meet some criteria. In contrast, our modified approach only performs iterative global refinement, which leads to significantly shorter computation times.

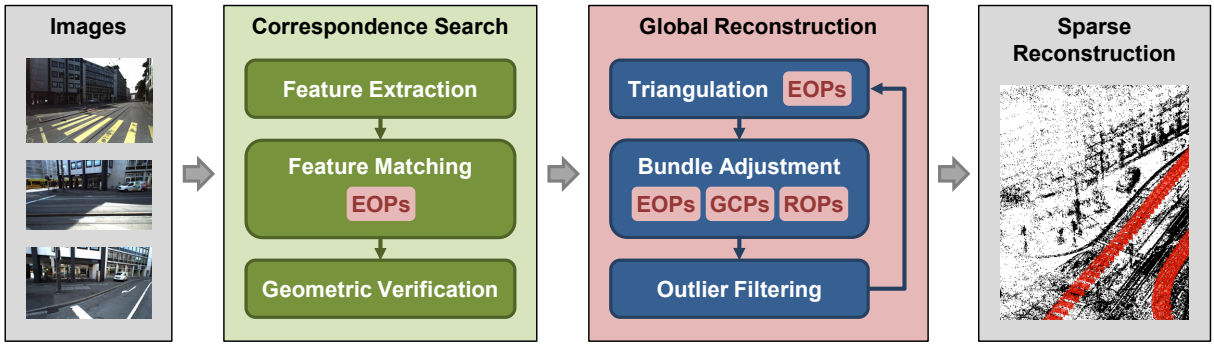

Figure 3.7: Adapted processing pipeline of COLMAP based on initial exterior orientation parameters (EOPs). Our contributions are marked in red, and they basically include a global SfM procedure exploiting EOPs, GCPs as well as ROPs (compare with figure 2.2).

Our integrated georeferencing approach utilizes three COLMAP modules, namely the *feature_extractor* and the *spatial_matcher* for correspondence search as well as the *mapper* module for global reconstruction (see figure 3.8). These modules contain the core functionality and they are implemented in C++. Moreover, we developed additional data preprocessing and post-processing procedures in Python, ranging from measuring pixel coordinates of GCPs to camera pose transformations.

3D coordinates of projection centers are given in a predefined coordinate reference frame, however, COLMAP requires translation vectors in a local system. Therefore, we first reduce initial projection centers to the EOP centroid, i.e. 3D coordinate translation into a local coordinate system. This step prevents calculation inaccuracies since large values can cause numeric instabilities. Second, we compute translation vectors $\mathbf{t}$ based on rotation matrices $\mathbf{R}_{CV}$ and projection centers $\mathbf{X}_0$ as follows:

$$\mathbf{t} = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^{\mathrm{T}} = -\mathbf{R}_{CV}\,\mathbf{X}_0 \tag{3.13}$$

(see section 7.1.2 for details on rotation matrices). Furthermore, we convert Euler angles as omega ω, phi φ, kappa κ to quaternions q_w, q_x, q_y, q_z (see section 7.1.2 for formulas). Correspondence search and global reconstruction result in refined local poses, which are transformed back to the predefined coordinate reference frame. Local projection center coordinates $\mathbf{X}_0$ are computed based on rotation matrices $\mathbf{R}_{CV}$ and translation vectors $\mathbf{t}$

$$\mathbf{X}_0 = \begin{bmatrix} X_0 & Y_0 & Z_0 \end{bmatrix}^{\mathrm{T}} = -\mathbf{R}_{CV}^{\mathrm{T}}\,\mathbf{t} \tag{3.14}$$

, and quaternions are converted to Euler angles (see section 7.1.2 for formulas).

We usually employ a few GCPs that are available in a predefined coordinate reference frame. Same as projection center coordinates, the 3D coordinates of these GCPs are reduced to the EOP centroid. These local coordinates do not require a further calculation, but can directly be utilized within the bundle adjustment process. ROPs are defined in a different local coordinate system. The reference camera defines its origin, the x-axis frequently points to the stereo partner, the y-axis upwards and the z-axis complements the right-handed coordinate system. In order to compute the desired components in a local system, equation (3.13) is needed that is the same as for EOPs.

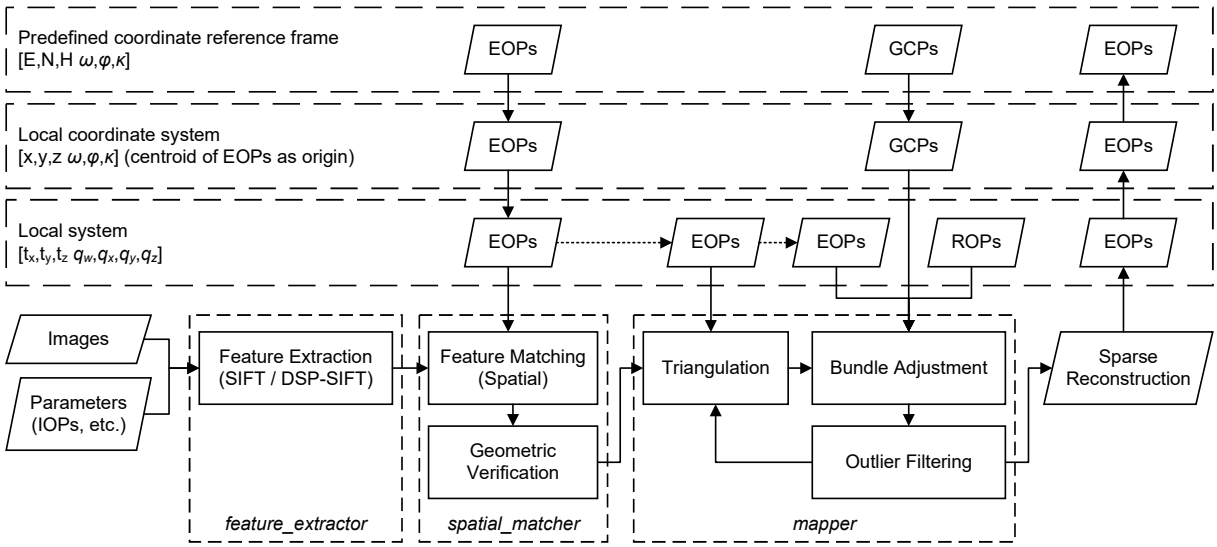

Figure 3.8: Adapted processing pipeline indicating *COLMAP modules* as well as required camera pose and 3D coordinate transformations.

Correspondence Search

As depicted by figure 3.7 and figure 3.8, correspondence search comprises feature extraction, feature matching and geometric verification. COLMAP extracts SIFT features by default. However, Schönberger et al. (2017) show that DSP-SIFT (Dong and Soatto, 2015) performs better than SIFT. Since more

features are extracted at the cost of longer computation times, DSP-SIFT is particularly advantageous in difficult environments with poor texture. We rely on initial EOPs, so that the spatial feature matcher implemented in COLMAP can be used. In order to reduce the search space, it only considers camera positions closer than a given maximum radius from the current image. Moreover, we added a maximum angle constraint to further speed up the process, as feature matching is the most time-consuming step within the COLMAP procedure.

Geometric verification of potentially overlapping image pairs is performed as described by Schönberger and Frahm (2016). SfM verifies the matches by trying to estimate a transformation that maps feature points between images using projective geometry. If a valid transformation maps a sufficient number of features between the images, they are considered as geometrically verified. Since the correspondences from matching are often outlier-contaminated, robust estimation techniques, such as RANSAC, are then required. The output is a scene graph with images as nodes and verified pairs of images as edges.

Global Reconstruction

Verified feature matches are triangulated to natural 3D points based on prior EOPs (see figure 3.9). Particularly in straight segments, same tie points are visible in numerous images. Hence, COLMAP tries to extend 3D point tracks, and merges 3D points that are very close to each other, followed by retriangulation. Bundle adjustment jointly refines camera poses and scene structure in a non-linear optimization (see section 2.1.5). We carry out global bundle adjustment based on Google's Ceres Solver library for non-linear least squares problems (Agarwal et al., 2020). Our bundle adjustment procedure (see equation (3.15)) minimizes reprojection errors between projected natural 3D points as well as ground control points and their corresponding 2D measurements in image space (see equation (3.16)). Moreover, same as Rumpler et al. (2017), it also minimizes differences of 3D projection center coordinates from direct georeferencing or SLAM and photogrammetric reconstruction (see equation (3.17)). We use either zero loss or the Cauchy loss function to potentially down-weight outliers. However, mobile mapping datasets frequently feature a moderate number of outliers, so that the robust Huber loss function would be an adequate option as well.

Optimization problem:

$$f^* = \min \sum E(P) + \sum E^{gcp}(R) + \sum E^{dg}(S) \tag{3.15}$$

where E = error function
P = natural 3D points
R = reference / ground control points
S = projection centers (from direct georeferencing or SLAM)

Error function for 3D points (tie points and GCPs):

$$E(X) = \sum_{x^i \in X_P} \rho(C^P(\Gamma_i(X), x^i)) \tag{3.16}$$

where ρ = loss function (e.g. robust Cauchy function)
C^P = 2D reprojection error / Euclidean distance in 2D
$\Gamma(X)$ = projected 3D point into image
x = observed 2D measurement

Error function for projection centers:

$$E(C) = \sum_{M \in S} \rho(C^{dg}(M, C)) \tag{3.17}$$

where ρ = loss function (e.g. robust Cauchy function)
C^{dg} = Euclidean distance in 3D
M = projection center from direct georeferencing or SLAM
C = reconstructed projection center

Besides, we enforce calibrated relative orientation parameters or define constraints for ROPs among cameras in bundle adjustment (e.g. constant ROPs for base over all image sequences if self-calibrated). COLMAP then completes and merges 3D point tracks, removes inconsistent points, retriangulates observations before performing a new bundle adjustment computation. This iterative global refinement process is continued until convergence is reached.

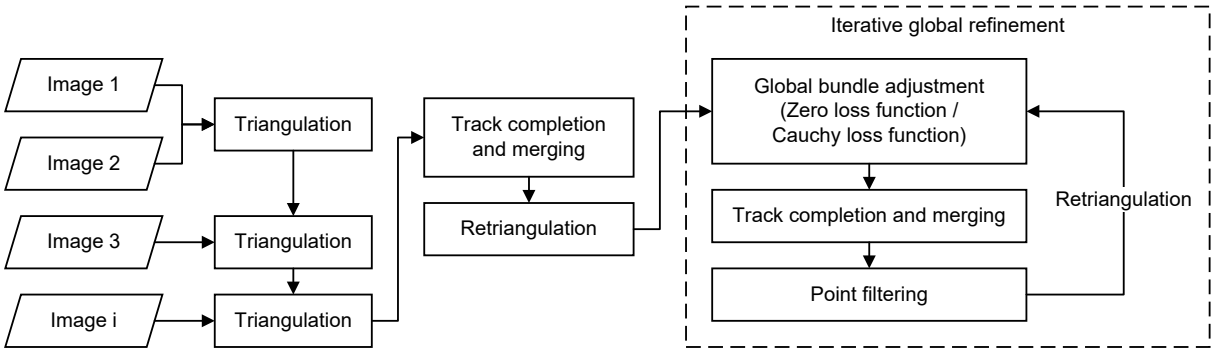

Figure 3.9: Detailed workflow of our global reconstruction procedure.

Summary

We presented our main contributions and their implementations in COLMAP within this chapter. Constraining relative orientation parameters among cameras allows for precise and robust image orientations, while incorporation of prior camera poses and ground control points leads to accurate georeferencing. In chapter 4, we evaluate our integrated georeferencing approach based on datasets featuring different environments and varying multi-camera systems.

Chapter 4

Evaluation of Integrated Georeferencing Approach

We aim at improving moderately accurate camera poses from direct georeferencing or SLAM in order to enable high-quality dense image matching. Since relying on multi-camera systems and exploiting relative orientation constraints (see section 3.1), image redundancy can be leveraged. For efficiency reasons, not the maximum but the most adequate number of cameras is of interest. Therefore, we evaluate varying camera configurations in different environments. Road and rail scenes often feature well textured areas that enable sufficient feature matching. In contrast, indoor environments are more challenging due to repetitive structures and weakly textured surfaces (see section 4.5). Furthermore, GNSS signals cannot be received indoors, so that a SLAM procedure is required for initial camera pose determination. We performed extensive investigations and show that our integrated georeferencing approach based on COLMAP is universally applicable. While GCP incorporation allows for high accuracies, solely employing EOPs increases efficiency.

4.1 Use Cases and Evaluation Methodology

We considered six datasets to determine the performance of our developed integrated georeferencing approach in terms of accuracy, robustness, efficiency and versatility. Three street-level datasets as well as one rail dataset were all captured by multi-stereo camera systems. While one stereovision system always pointed forward, additional stereo camera systems enabled a close to 360° horizontal coverage. A full horizontal coverage was reached in our two indoor use cases by utilization of a multi-head panorama camera. Such configurations enable strong feature point connections that lead to accurate and robust results.

4.1.1 Overview of Test Campaigns

We evaluate our integrated georeferencing approach using several real-world datasets (see table 4.1). They comprise road, rail and indoor environments. While prior EOPs were obtained by direct georeferencing outdoors, lidar SLAM delivered initial EOPs in GNSS denied environments such as buildings (see section 3.3.2). The two datasets Basel14 and Basel15 cover the same study area in the city center of Basel (see section 4.2). It includes a junction with three street sections. Nonetheless, the utilized camera configurations vary considerably. While three horizontally arranged stereo camera systems served for Basel14, Basel15 features five stereo systems and two further cameras. Apart from a forward pointing stereo camera system, there are four stereo systems constituted by panorama camera heads that enable an adequate vertical coverage. A standard stereo Y configuration featuring stereo camera systems facing forward, back-right and back-left was employed for capturing the dataset Zug17 (see section 4.3). It includes five street loops in a suburban environment. The rail dataset Vienna16 involves a train station and was acquired by five stereo camera systems (see section 4.4). In contrast, a panorama camera collected

the indoor datasets Muttenz17 and Muttenz18 that cover the same floor of a building (see section 4.5). Indoor environments typically show homogeneous surfaces and repetitive patterns, which pose a major challenge for feature matching.

Name	Month	Environment	Prior EOPs	Used camera config.	# Cam.
Basel14	07.2014	road: urban (junction)	GNSS/INS (DG)	2x forward, 2x back-right, 2x left	6
Basel15	08.2015	road: urban (junction)	GNSS/INS (DG)	2x forward, 2x tilted panorama (2x 5)	12
Zug17	03.2017	road: suburban (five loops)	GNSS/INS (DG)	standard stereo Y: 2x forward, 2x back-right, 2x back-left	6
Vienna16	10.2016	rail: train station	GNSS/INS (DG)	4x forward, 2x downward, 2x right, 2x left	10
Muttenz17	11.2017	indoor: building	Lidar SLAM	1x horizontal panorama (5 camera heads)	5
Muttenz18	03.2018	indoor: building	Lidar SLAM	1x horizontal panorama (5 camera heads)	5

Table 4.1: Overview of datasets and utilized multi-camera systems. Road and rail environments were mapped by multi-stereo systems, while a panorama camera served for indoor mapping.

Our integrated georeferencing approach relies on initial EOPs, exploits ROPs among cameras, and can employ GCPs. Varying environments lead to different accuracies of initial EOPs. Its influence as well as the benefit of not only using a forward pointing stereo camera system are extensively investigated within the next sections. According to table 3.1, we differentiate four use cases by utilization of GCPs or not, and either by fixing ROPs or by self-calibrating them. Due to GCP incorporation, use cases I and II allow for highest accuracies. Moreover, exploitation of ROPs among cameras attached to a rigid frame greatly increases the chance that all images can be oriented, i.e. high robustness. As shown by table 4.2, we fixed precalibrated ROPs for our indoor datasets, while self-calibrating ROPs in the outdoor cases. Determination of GCP coordinates is a costly task. Hence, we demonstrate the accuracy potential of integrated georeferencing without GCPs for an extended junction in an urban environment as well as for several loops in a suburban neighborhood. Moreover, we show that a multi-camera configuration of a train-based MMS can be self-calibrated by only fixing stereo camera bases.

Name	I	II	III	IV	V
Basel14		4.2.2, 4.2.3			5
Basel15		4.2.2, 4.2.3	4.2.4		
Zug17		4.3.2	4.3.3		
Vienna16				4.4.2	
Muttenz17	4.5.2				
Muttenz18	4.5.2				

Table 4.2: Chapter indication of datasets and respective use cases defined in table 3.1 for evaluation of our integrated georeferencing approach. I: IG with GCPs and fixed ROPs, II: IG with GCPs and ROP self-calibration, III: IG without GCPs but with fixed ROPs, IV: IG without GCPs but with ROP self-calibration, V: In-sequence dense image matching.

4.1.2 Standard Processing and Investigation Procedure

One of our main features is the exploitation of relative orientation constraints (see section 3.1). Hence, the ROP configuration and its corresponding calibrated values need to be defined in a JSON file as expected from COLMAP. For all of our outdoor use cases, the left camera of the forward pointing stereo system

serves as reference of the multi-stereo camera rig. Indoors is different, since the backward facing camera head of the panorama camera acts as reference.

We mainly used the predefined COLMAP standard parameters. Therefore, SIFT features were extracted for all of our outdoor datasets. However, we performed DSP-SIFT feature extraction for our indoor datasets Muttenz17 and Muttenz18. While DSP-SIFT delivers more features that is especially beneficial in difficult environments with weakly textured areas, computation time is significantly increased (see section 3.4). We frequently used a maximum angle constraint of 100 degrees within the spatial feature matching procedure, which would even allow for matching of images from forward and left pointing systems. Although we usually utilized a maximum radius of 20 m, we used 10 m for the datasets Basel14 and Basel15 when processed with GCPs as well as for our indoor datasets. We did not refine any interior orientation parameters within the bundle adjustment process, since they were previously calibrated precisely (see section 3.3.1). Empirical experiments showed that the linear solver type SPARSE_SCHUR needs considerably less computation time than ITERATIVE_SCHUR for bundle adjustment due to faster convergence. Furthermore, we considered a maximum of ca. 10'000 images within a single processing, which is still feasible with SPARSE_SCHUR and thus our prime choice. We utilized the zero loss function due to a rather low amount of outliers and in order to avoid down-weighting of GCP observations. Nonetheless, investigations without GCPs for Basel15 and Zug17 were performed using the robust Cauchy loss function.

In order to determine the achievable mapping accuracy, comparing 3D coordinates of check points with a more accurate reference is the standard procedure. This reference is frequently captured by tachymetry. Check point coordinates are computed by spatial intersection employing exterior orientation parameters from integrated georeferencing as well as image point measurements performed manually. For the datasets Basel14 and Basel15, we used a Matlab tool and determined 3D point coordinates based on single stereo image pairs from the forward pointing camera system, i.e. only two image observations for each point were provided. For the other datasets, we performed a bundle adjustment-based forward intersection with a Python tool usually employing four image observations per point.

4.2 Test Campaigns in an Urban Environment

Urban environments frequently show various scene structures and distinctive textures, which is beneficial for feature extraction and matching. However, prevalent large buildings in cities often form urban canyons. These pose a challenge for GNSS positioning due to GNSS outages and multipath effects. Hence, direct georeferencing accuracies of several decimeters up to meters are the normal case in such environments. We performed two road mapping campaigns using different camera configurations in the city center of Basel, Switzerland. The first configuration includes three stereo camera systems, while the other incorporates twelve cameras acquiring highly redundant imagery. In a first set of experiments, we assessed the quality of directly georeferenced image orientations as well as its improvement by integrated georeferencing. Subsequently, we compared different scenarios regarding camera configurations and GCP groups. Finally, the potential of integrated georeferencing without employing any GCPs was investigated.

4.2.1 Vehicle-Based Mobile Mapping Systems and Data

Mobile Mapping Platform

The vehicle-based stereovision mobile mapping research platform of the Institute of Geomatics at FHNW features several industrial stereo cameras with CCD sensors as well as a GNSS/IMU positioning system (see figure 3.1 and figure 4.1). All sensors are synchronized by hardware trigger signals from a custom-built trigger box. Although the platform allows for arbitrary camera configurations, there is always a main stereovision system with high-resolution cameras facing forward covering the road with its infrastructure (see table 4.3 and table 4.4). For the campaign in July 2014, we additionally assembled FHD cameras forming a back-right stereovision system mapping the closer sidewalk area and lower façade parts as well as a left stereovision system covering the opposite sidewalk area. However, we did not consider imagery captured by a panorama camera in our experiments. For the campaign in August 2015, we used our novel 360° stereo panorama setup with two multi-head panoramic cameras equipped with fisheye optics tilted

forward and backward by 90° each (Blaser et al., 2018; Nebiker, 2019). In addition to one single camera head facing forward and one backward, the other individual heads of the panoramic cameras pointing sideways constitute five stereo systems. They cover pavement, complete façades of buildings and the entire overhead space even in heavily built-up urban environments. However, we disregarded imagery from the stereo system looking downward at the mobile mapping vehicle as well as images captured by an additional FHD camera in the center of the forward pointing stereo camera system (see figure 4.2).

Our MMS features a NovAtel SPAN inertial navigation system for direct georeferencing of the imagery acquired at typically 5 fps. The navigation system consists of a tactical grade inertial measurement unit featuring fiber-optics gyros of the type UIMU-LCI and a L1/L2 GNSS kinematic antenna. In case of good GNSS coverage, these sensors provide an accuracy of horizontally 10 mm and vertically 15 mm during post-processing. Accuracies of the attitude angles roll γ and pitch θ are specified with 0.005° and heading ψ with 0.008°. A GNSS outage of 60 seconds degrades the horizontal accuracy to 110 mm and the vertical to 30 mm.

Figure 4.1: Multi-view multi-sensor stereovision IGEO mobile mapping system with sensor configuration for the campaign in July 2014 (left) and for the campaign in August 2015 (right).

Camera type	Sensor size	Pixel size [µm]	Principal distance [mm]	Field of view [°]	Camera model
AVT	11 MP (4008 × 2672)	9.00	21.0	81 × 60	perspective
Basler/FHD	2 MP (1920 × 1080)	7.40	7.9	84 × 54	perspective
Ladybug 5 camera head	5 MP (2448 × 2048)	3.45	4.3	113 × 94	fisheye

Table 4.3: Interior orientation parameters of the three different camera types mounted on the IGEO mobile mapping system.

	Basel14			Basel15	
Pointing direction of stereo cameras	forward	back-right	left	forward	left-down, left-up, right-up, right-down
Camera type	AVT	Basler	Basler	AVT	Ladybug 5
Base length [mm]	905	779	949	905	1584

Table 4.4: Length of stereo bases for the two camera configurations used during the campaigns in Basel.

Figure 4.2: Mobile mapping images captured by all cameras at the same location during the campaign in August 2015 (forward stereo [top left], panorama forward and backward [top right], panorama stereo [middle and bottom]).

Study Area and Mobile Mapping Data

The relatively small but demanding study area depicted in figure 4.3 is located at a very busy junction of five roads in the city center of Basel, Switzerland. It includes three streetcar stops leading to many overhead wires as well as large and rather tall commercial properties (see figure 4.4) that create a very challenging environment for GNSS positioning (see figure 4.6). Besides, there is ongoing construction work as well as several moving objects such as pedestrians, cars and streetcars in the investigated region, which pose even more challenges for data processing. Three street sections of this test site featuring sidewalks were mapped three times, once in July 2014 and twice during a day in August 2015, which is a difference in time of 13 months (see table 4.5). In all nine cases data acquisition was performed shortly before noon and in good weather conditions. The individual image sequences contain 85 up to 191 timestamps on a sequence length between 108 m and 217 m. An along-track distance between successive image exposures of 1 m was targeted, but larger distances occurred at velocities higher than 18 km/h since the maximum frame rate was 5 fps.

While the campaign in July 2014 was part of a complete survey of the city-state of Basel, the campaign in August 2015 was specifically performed for the investigations at our study area (see figure 4.5). In order to capture optimal trajectories, we collected kinematic data according to best practice as specified by the manufacturer. First, static initialization for approx. three minutes in an open sky area followed by leveling until approaching the test site was carried out. After the first mapping of the test site, an additional loop was driven so that data could again be acquired in the study area. Returning to the start area, imagery was captured on our outdoor calibration field for the purpose of boresight alignment. A further loop served for leveling and there was a static observation at the end of around four minutes nearby the former FHNW building as well. A GNSS station on its roof, which used to be part of the Automated GNSS Network for Switzerland (AGNES), was defined as base station. The complete campaign resulted in a total trajectory length of 22.756 km and 12'220 image epochs captured on 20.8.2015 from 10:17:53 until 11:19:29. Section 3.3 covers calibration and direct georeferencing for the dataset Basel14, while Blaser et al. (2018) describe in detail the complete calibration and processing procedure for the dataset Basel15.

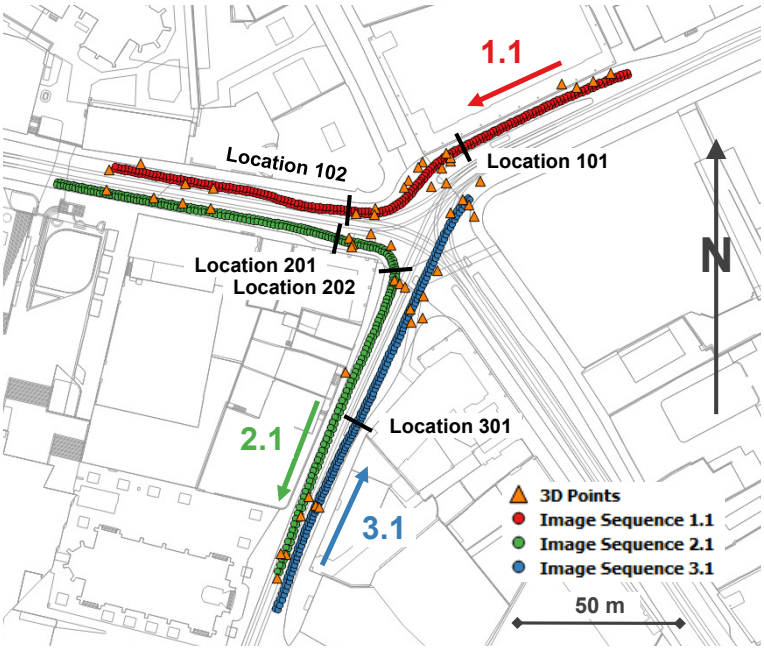

Figure 4.3: Base map of the study area with overlaid projection centers of selected stereo image sequences, 3D reference points and locations of trajectory discontinuities (Source of base map: Geodaten Kanton Basel-Stadt).

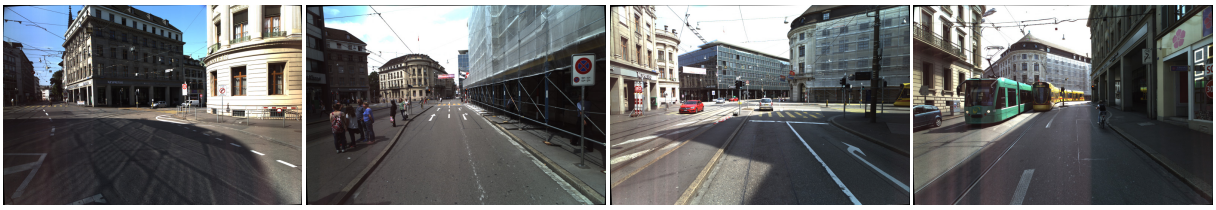

Figure 4.4: Mobile mapping imagery of the campaign performed in July 2014 illustrating the test site characteristics but also typical challenges, e.g. GNSS shadowing, numerous pedestrians, heavy traffic with multiple streetcars, cars and cyclists.

Sequence	Date and Time	# Epochs	Length [m]	Along-track spacing [m]	
				Mean	Max.
1.0	24.7.14 10:20	123	164	1.34	1.97
1.1	20.8.15 10:30	161	173	1.08	1.25
1.2	20.8.15 10:47	156	175	1.13	1.48
2.0	27.7.14 11:53	157	173	1.11	1.60
2.1	20.8.15 10:34	171	212	1.25	2.06
2.2	20.8.15 10:50	191	217	1.14	1.93
3.0	27.7.14 11:57	85	108	1.29	1.49
3.1	20.8.15 10:37	116	141	1.23	1.73
3.2	20.8.15 10:53	96	146	1.54	2.37

Table 4.5: Characteristics of the nine selected image sequences x.y (where x corresponds to the street sections 1 to 3 shown in figure 4.3 and y indicates the campaign, 0 = 24./27.7.2014, 1 = 20.8.2015 10:30-10:37, 2 = 20.8.2015 10:47-10:53).

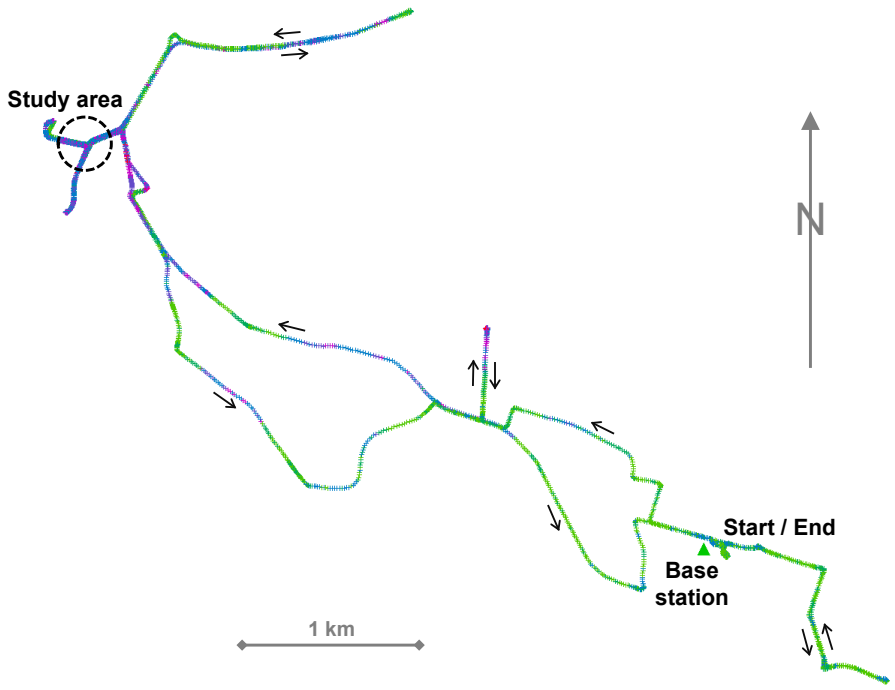

Figure 4.5: Trajectory of the campaign performed on 20.8.2015 (green: high direct georeferencing quality, red: low quality, study area: medium to low quality; trajectory extent in east-west direction is approximately 4.75 km).

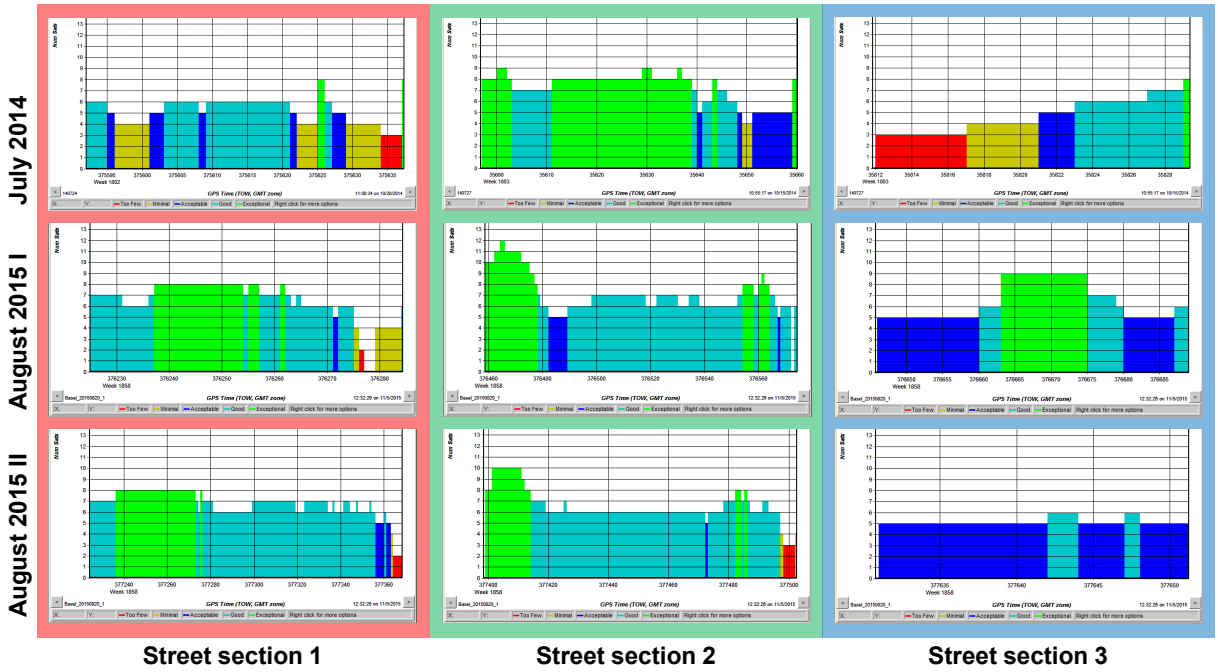

Figure 4.6: Number of satellites for all image sequences captured in Basel (blue equals five satellites). The variations for each street section are significant, even though campaign 2 was only performed 16 minutes later than campaign 1 (compare middle with bottom row).

Reference Data

In order to validate our results, we captured independent and highly accurate reference data in March 2015, which is eight months after the first mobile mapping survey. Nonetheless, there were no significant changes for permanent objects such as buildings and roads in the study area. However, changes occurred due to moving objects.

We performed four 360° terrestrial laser scans recording XYZ point geometry and intensity (see figure 5.5). By registering the point clouds onto several cadastral reference points, an absolute 3D TLS accuracy of 1-2 cm was obtained. In addition, we determined 3D coordinates of more than 50 points mainly on corners of road markings using a total station. These points have an absolute 3D accuracy of better than 1 cm and served either as ground control or check points. All measurements were performed in the Swiss horizontal reference frame LV95 and with leveled heights in the LN02 vertical reference frame.

4.2.2 Significant Improvement of Direct Georeferencing Solution by Integrated Georeferencing

By performing both trajectory and check point investigations, we show within this section that direct georeferencing solutions consistently deviate several decimeters in challenging urban environments. However, these offsets as well as trajectory discontinuities can be corrected by integrated georeferencing.

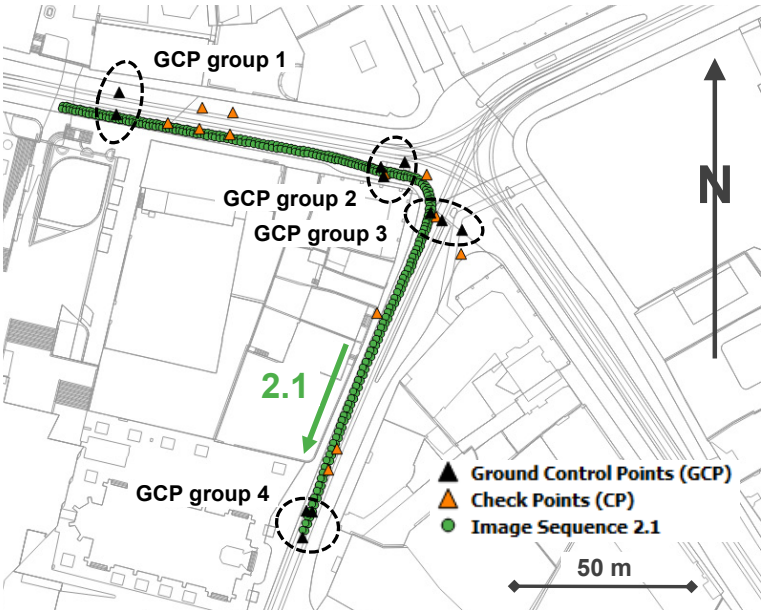

Figure 4.7: Locations of ground control point groups as well as check points for stereo image sequence 2.1 (Source of background map: Geodaten Kanton Basel-Stadt).

Data Processing

We established one GCP group consisting of two, three or four GCPs at each end of a street section as well as two additional GCP groups in-between close to the corresponding sharp curve (see figure 4.7). Since street section 3 is a straight segment, it was defined by only two GCP groups. Identification and sensor coordinate measurement of the utilized natural GCPs mainly on corners of lane markings or crosswalks was sometimes challenging. We performed integrated georeferencing with ROP self-calibration for each of the nine forward stereo image sequences by incorporating exterior orientation parameters from direct georeferencing. Hence, ROPs of the forward pointing stereo base were constrained but not fixed. This resulted in a mean track length, which is the mean number of images that observe the same tie point, of

at least 7 for all sequences (see table 4.6). Furthermore, all values for mean observations per image are larger than 3000 and mean reprojection errors between 0.71 and 0.83 pixel were computed. Please note that no separation between GCPs and tie points but the total values are depicted.

Sequence	1.0	1.1	1.2	Mean 1.x
Registered images	246	322	312	
3D points	126'568	160'896	150'814	
Observations	885'720	1'191'775	1'052'595	
Mean track length	7.0	7.4	7.0	7.1
Mean obs. per image	3601	3701	3374	3559
Mean reproj. error [px]	0.73	0.76	0.81	0.77
Sequence	2.0	2.1	2.2	Mean 2.x
Registered images	314	342	382	
3D points	156'445	154'529	166'459	
Observations	1'200'287	1'259'882	1'308'171	
Mean track length	7.7	8.2	7.9	7.9
Mean obs. per image	3823	3684	3425	3644
Mean reproj. error [px]	0.71	0.79	0.81	0.77
Sequence	3.0	3.1	3.2	Mean 3.x
Registered images	170	232	192	
3D points	90'498	107'511	83'328	
Observations	708'763	832'515	583'823	
Mean track length	7.8	7.7	7.0	7.5
Mean obs. per image	4169	3588	3041	3599
Mean reproj. error [px]	0.74	0.83	0.81	0.79

Table 4.6: COLMAP processing statistics of datasets Basel14 and Basel15. Four GCP groups were considered for street sections 1 and 2, while two GCP groups were incorporated for street section 3.

Trajectory and Orientation Deviations between Direct and Integrated Georeferencing

In order to assess the quality of directly georeferenced sensor orientations as well as the potential improvement by integrated georeferencing in a challenging urban environment with frequent GNSS degradations, we computed deviations of projection centers and orientation angles between direct and integrated georeferencing for all nine sequences. 3D deviations of projection centers range from 59 mm to 804 mm and result in a mean value of 400 mm (see table 4.7 and figure 4.8). The height is the component with the largest residuals for all but for sequences 3.1 and 3.2. We obtained rather small deviations for street section 3, for both projection centers and orientation angles (see table 4.8). While sequences 1.1 and 2.1 show the largest omega ω and kappa κ deviations with more than 0.4°, we achieved mean values of ca. 0.25° for these components. Phi φ values are rather small for all sequences.

Sequence	ΔE [mm]	ΔN [mm]	ΔH [mm]	Δ3D [mm]
1.0	297	38	425	520
1.1	38	89	125	159
1.2	436	29	568	716
2.0	54	40	83	107
2.1	266	92	502	576
2.2	173	478	624	**804**
3.0	31	17	88	94
3.1	29	42	30	**59**
3.2	166	520	135	562
Mean	166	149	287	**400**

Table 4.7: RMSE values for deviations of projection centers between direct and integrated georeferencing.

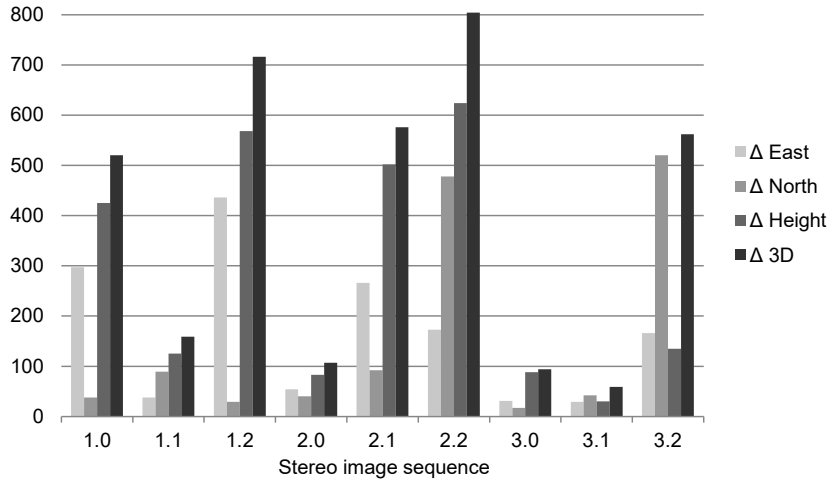

Figure 4.8: RMSE values in mm for deviations of projection centers between direct and integrated georeferencing depicted in a chart (same values as in table 4.7).

Sequence	$\Delta\omega$ [°]	$\Delta\varphi$ [°]	$\Delta\kappa$ [°]
1.0	0.240	0.090	0.239
1.1	**0.442**	0.068	**0.418**
1.2	0.389	0.071	0.366
2.0	0.069	0.049	0.080
2.1	**0.405**	0.080	**0.440**
2.2	0.381	0.062	0.431
3.0	0.023	0.058	0.038
3.1	0.103	0.087	0.086
3.2	0.094	0.063	0.143
Mean	**0.238**	**0.070**	**0.249**

Table 4.8: RMSE values for deviations of image orientation angles between direct and integrated georeferencing.

Potential improvements in deviations of projection centers and thus in trajectory accuracy due to integrated georeferencing over direct georeferencing are illustrated in detail by figure 4.9 to figure 4.17. Trajectories of stereo image sequences captured on the same street section at different times show differences of up to several decimeters. While we obtained small deviations for sequences 1.1 and 2.0, they are

significantly larger for the other sequences of these two street sections. All deviations of street section 3 are smaller than 20 cm, with the exception of the north component of sequence 3.2, which amounts to approx. 50 cm. All sequences of street section 1 basically show positive deviations in the east component, negative height deviations, and north deviations that are close to zero.

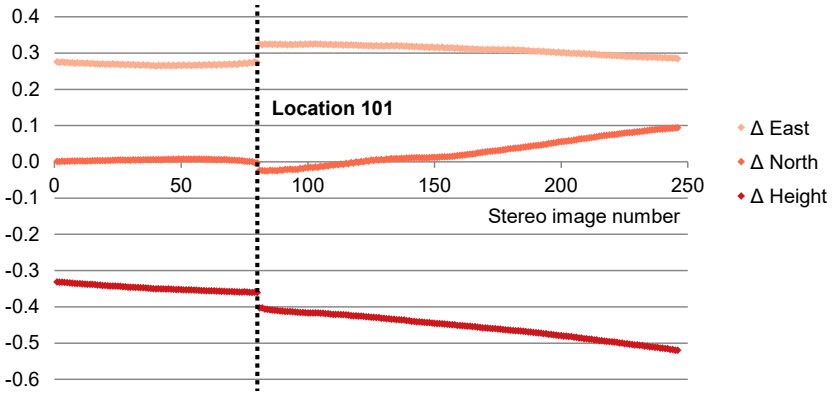

Figure 4.9: Deviations of projection centers in m for image sequence 1.0. Residuals of the east component as well as height deviations amount to several decimeters. There is a trajectory discontinuity at location 101.

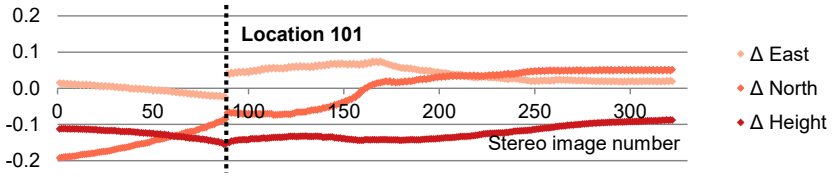

Figure 4.10: Deviations of projection centers in m for image sequence 1.1. All differences are smaller than 20 cm, but there is a trajectory discontinuity at location 101.

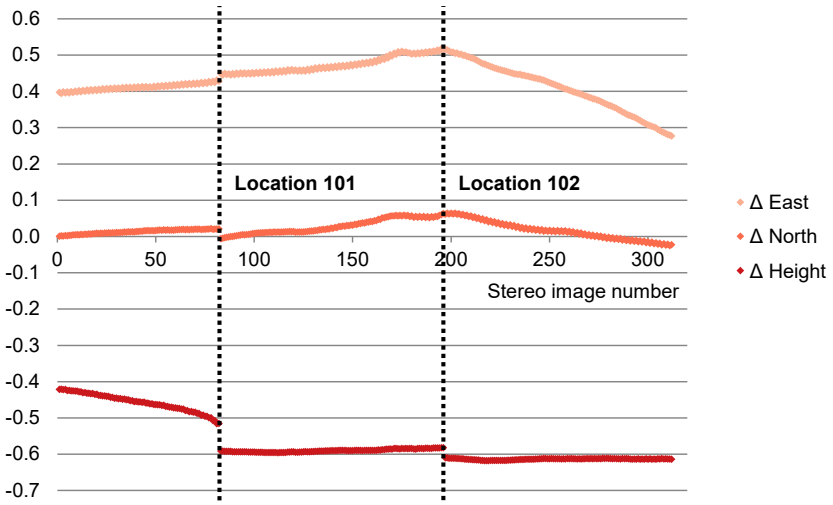

Figure 4.11: Deviations of projection centers in m for image sequence 1.2. Large residuals for both the east component and the height were obtained. Moreover, the computed deviations disclose two trajectory discontinuities, one at location 101 and one at location 102.

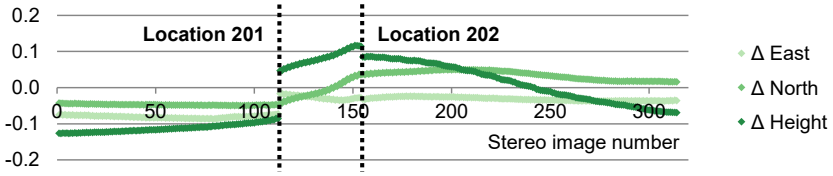

Figure 4.12: Deviations of projection centers in m for image sequence 2.0. Even though differences are small, they reveal two trajectory discontinuities at locations 201 and 202.

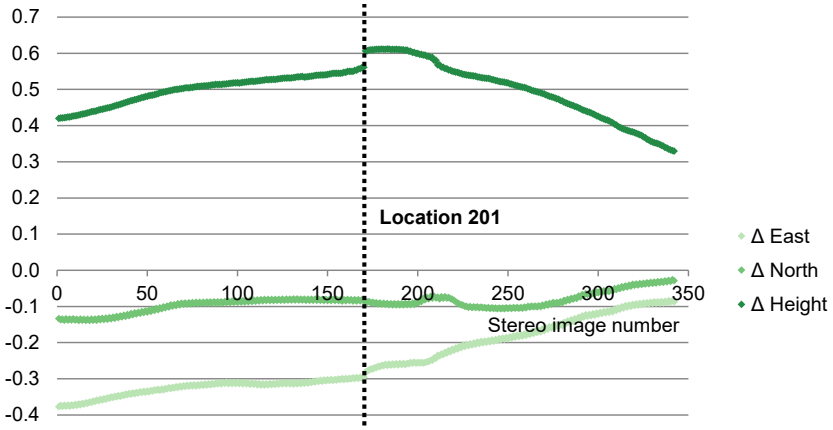

Figure 4.13: Deviations of projection centers in m for image sequence 2.1. The height component shows residuals of more than 30 cm as well as a significant discontinuity at location 201.

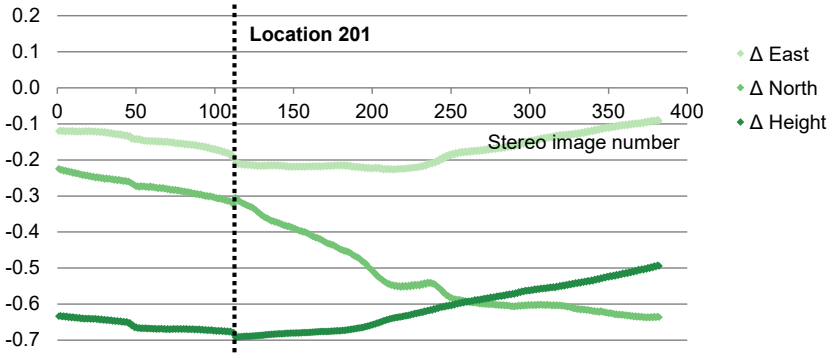

Figure 4.14: Deviations of projection centers in m for image sequence 2.2. All differences are negative and they range up to 70 cm. There is a trajectory discontinuity at location 201.

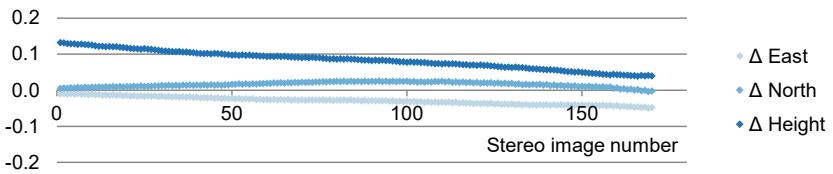

Figure 4.15: Deviations of projection centers in m for image sequence 3.0. Most differences are smaller than 10 cm.

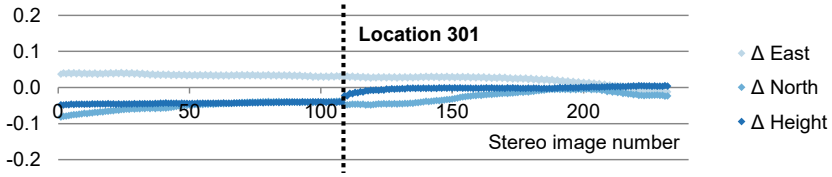

Figure 4.16: Deviations of projection centers in m for image sequence 3.1. Although all differences lie within the sub-decimeter range, they reveal a trajectory discontinuity at location 301.

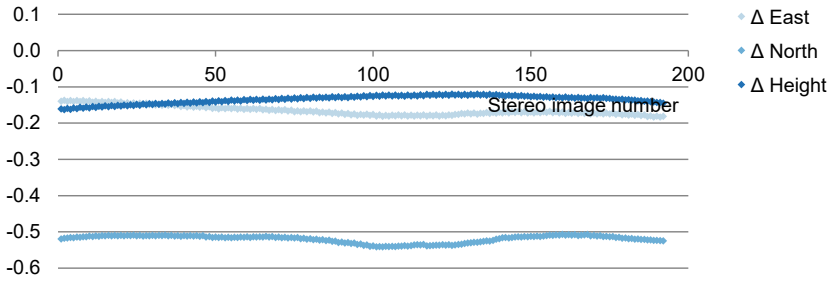

Figure 4.17: Deviations of projection centers in m for image sequence 3.2. The north component deviates by ca. 50 cm, while the other two components show differences of less than 20 cm.

Trajectory Discontinuities in Direct Georeferencing

The charts illustrating trajectory deviations between direct and integrated georeferencing clearly reveal nine trajectory discontinuities, which are indicated by vertical dotted lines (see figure 4.9 until figure 4.17). According to figure 4.3 and figure 4.18 all of them but one (location 202) were caused by a vehicle stop of several seconds mainly in front of crosswalks. However, no correlation between stop duration and 3D value of the discontinuities could be proven (see table 4.9). 3D discontinuities amount mostly to a few centimeters, but they reach up to approx. 15 cm for sequence 2.0 at location 201. No discontinuities are present in sequences 3.0 and 3.2, mainly because of no vehicle stops. There is an option for fine-tuning the automated zero velocity update (ZUPT) detection tolerances in the GNSS/INS post-processing software Inertial Explorer, which might eliminate trajectory discontinuities, but not remove the observed systematic trajectory offsets.

Location	Sequence	Stop [s]	ΔE [mm]	ΔN [mm]	ΔH [mm]	Δ3D [mm]
101	1.0	17	49	-21	-43	68
101	1.1	13	63	19	4	66
101	1.2	31	17	-26	-75	81
102	1.2	68	-1	1	-28	28
201	2.0	19	56	4	133	144
201	2.1	56	16	-2	46	49
201	2.2	41	-15	15	-12	24
202	2.0	0	-4	1	-30	30
301	3.1	16	-2	-11	18	21

Table 4.9: Dimensions of trajectory discontinuities that we disclosed at nine locations.

Figure 4.18: Mobile mapping imagery captured at locations of trajectory discontinuities.

Check Point Investigations for Direct and Integrated Georeferencing

[mm]	Seq.	# CPs	ΔE	ΔN	ΔH	Δ2D	Δ3D
IG	1.0	14	21	14	7	25	26
	2.0	10	10	15	6	18	19
	3.0	6	21	26	12	33	35
	Mean x.0		17	18	**8**	25	**27**
DG	1.0	14	318	59	445	324	551
	2.0	10	65	49	97	82	126
	3.0	6	47	20	76	51	92
	Mean x.0		144	43	**206**	152	**256**
Improvement	1.0	14	297	45	438	299	525
	2.0	10	55	34	91	64	107
	3.0	6	26	-5	64	18	57
	Mean x.0		127	25	198	127	229
	Factor DG		*8.5*	*2.4*	*25.8*	*6.1*	***9.5***

Table 4.10: RMSE values in mm for check point residuals of direct (DG) and integrated georeferencing (IG) in case of dataset Basel14.

We computed 3D check point coordinates based on EOPs from both direct and integrated georeferencing, and calculated deviations to reference 3D coordinates determined by tachymetry. The resulting

RMSE values for the dataset Basel14 are depicted in table 4.10. While there is a mean 3D RMSE value of 256 mm for direct georeferencing, it amounts to 27 mm for integrated georeferencing, which is an improvement by an order of magnitude. This is mainly caused by a significant height improvement from 206 mm to 8 mm. As expected, all 3D RMSE values computed from check point residuals of direct georeferencing are similar to corresponding 3D RMSE values from trajectory deviations (compare table 4.10 with table 4.7).

4.2.3 Further Investigations on Exploiting Integrated Georeferencing with Ground Control Points and ROP Self-Calibration

Integrated georeferencing with ROP self-calibration enables high accuracies. Compared to single forward stereo, processing multi-stereo images in a road junction area allows to especially improve the height component. A similar effect can be observed if curved road segments are supported with additional GCPs.

Data Processing

Same as in section 4.2.2, we defined one GCP group at each end of a segment, but no GCP groups close to the sharp curves. We did also estimate ROPs among respective cameras over all stereo images. This corresponds to a self-calibration of stereo bases as well as of position vectors and rotations among the stereo camera systems. First, we processed images from the forward facing stereo system (single stereo) for all sequences of the dataset Basel14 as well as for sequences of street sections 1 and 2 of the dataset Basel15. Second, we performed integrated georeferencing exploiting all images from the three stereo camera systems (multi-stereo) of the dataset Basel14 within the same process (see figure 4.19). Shorter tie point tracks were computed for multi-stereo compared to single stereo (see table 4.11), however, multi-stereo has a better mean reprojection error of 0.51 pixel.

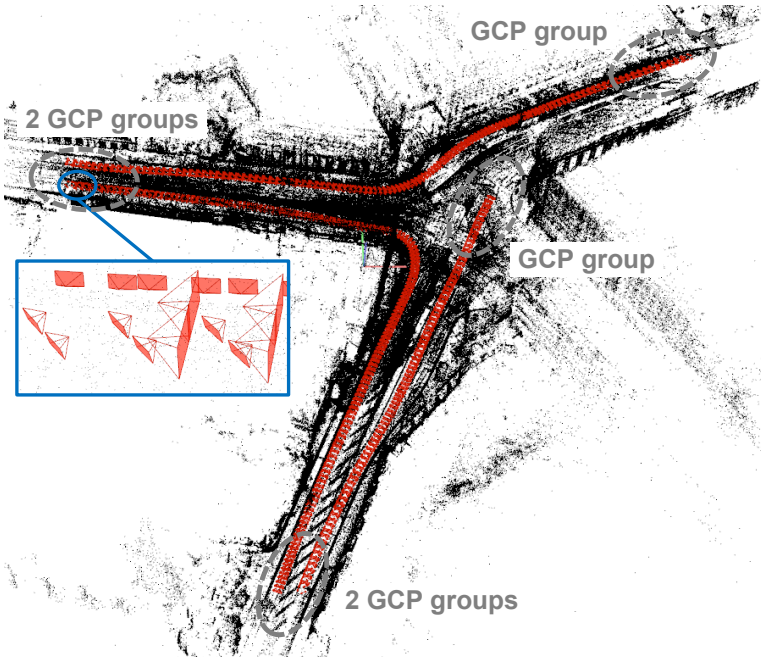

Figure 4.19: Georeferenced mobile mapping images (red) and 3D tie points (black) at our test site in Basel processed by our modified COLMAP procedure.

Please note that the results of table 4.11 do not exactly correspond to the results of table 4.6. Reasons are two instead of four GCP groups for street sections 1 and 2, but also a slightly modified processing workflow. While sequence 3.0 has fewer 3D points and observations in table 4.11, which leads to a smaller

value for mean observations per image, it features longer tie point tracks. Mean reprojection errors are similar for both processings.

	Single stereo 1.0	Single stereo 2.0	Single stereo 3.0	Multi-stereo
Registered images	246	314	170	2190
3D points	106'814	129'125	71'081	1'038'228
Observations	769'117	1'020'655	601'128	6'972'788
Mean track length	7.2	7.9	8.5	6.7
Mean obs. per image	3127	3251	3536	3184
Mean reproj. error [px]	0.72	0.70	0.72	0.51

Table 4.11: COLMAP processing statistics for single stereo and multi-stereo image sequences. Two GCP groups were incorporated for every street section.

Check Point Investigations for Single Stereo and Multi-Stereo

In order to verify the assumed advantages of using multi-stereo camera systems compared to single stereo, we computed check point residuals to tachymetry. As depicted by table 4.12, there is a mean 3D RMSE value of 39 mm for single stereo and a value of 28 mm for multi-stereo image sequences (combination of forward stereo, back-right stereo and left stereo). The mean height RMSE value reduces from 18 mm to 8 mm. This improvement is especially caused by image sequence 2.0, which shows a significant accuracy increase in the height component from 34 mm to 7 mm as well as in 3D from 51 mm to 22 mm. Results of multi-stereo processing employing GCP groups at each end of a segment are similar to results of single stereo processing that additionally incorporated GCP groups close to sharp curves (compare table 4.12 with table 4.10).

[mm]	Seq.	# CPs	ΔE	ΔN	ΔH	Δ2D	Δ3D
Multi-stereo	1.0	14	20	19	10	28	30
	2.0	10	13	16	**7**	20	**22**
	3.0	6	19	23	7	30	31
	Mean x.0		17	19	**8**	26	**28**
Single stereo	1.0	14	24	14	8	28	29
	2.0	10	26	28	**34**	38	**51**
	3.0	6	21	27	11	34	36
	Mean x.0		24	23	**18**	33	**39**
Improvement	1.0	14	4	-5	-2	0	-1
	2.0	10	13	12	27	18	29
	3.0	6	2	4	4	4	5
	Mean x.0		7	4	10	7	11
	Factor		*1.4*	*1.2*	*2.3*	*1.3*	*1.4*

Table 4.12: RMSE values in mm for check point residuals between integrated georeferencing and tachymetry regarding both single stereo and multi-stereo camera configurations.

Check Point Investigations for Two and Four GCP Groups

While we utilized four GCP groups on street sections 1 and 2 in chapter 4.2.2, there were only two GCPs groups per image sequence in chapter 4.2.3. The results of these two scenarios but also direct georeferencing values are illustrated in table 4.13. For scenario I with two GCP groups, we obtained a mean 3D RMSE value of 62 mm which is eight times better than a value of ca. 50 cm for direct georeferencing. Scenario II featuring four GCP groups, which led to 3D check point residuals per sequence of 19-56 mm and a mean 3D RMSE value of 33 mm, improves the direct georeferencing solution by a factor of 15. Scenario II shows a mean 3D improvement by a factor of two an a mean height improvement

by a factor of five when compared to scenario I. Stereo image sequences 2.1 and 2.2 have the largest deviations but also the largest improvements.

[mm]	Seq.	# CPs	ΔE	ΔN	ΔH	Δ2D	Δ3D	Δ3D DG
IG 4 GCP groups (scenario II)	1.0	14	21	14	7	25	26	551
	1.1	11	18	24	8	30	31	168
	1.2	10	20	14	5	24	25	764
	2.0	10	10	15	6	18	19	126
	2.1	11	27	27	8	39	40	567
	2.2	10	39	39	7	55	56	816
	Mean		23	22	**7**	32	**33**	499
	Mean DG		227	125	403	291	499	
	Factor DG		*9.9*	*5.7*	*57.6*	*9.1*	***15.1***	
IG 2 GCP groups (scenario I)	1.0	14	24	14	8	28	29	551
	1.1	11	18	21	11	28	30	168
	1.2	10	18	14	34	23	41	764
	2.0	10	26	28	34	38	51	126
	2.1	11	57	59	36	82	90	567
	2.2	10	79	56	89	96	131	816
	Mean		37	32	**35**	49	**62**	499
	Mean DG		227	125	403	291	499	
	Factor DG		*6.1*	*3.9*	*11.5*	*5.9*	***8.0***	
Improvement	1.0	14	3	0	1	3	3	
	1.1	11	0	-3	3	-2	-1	
	1.2	10	-2	0	29	-1	16	
	2.0	10	16	13	28	20	32	
	2.1	11	30	32	28	43	50	
	2.2	10	40	17	82	41	75	
	Mean		14	10	28	17	29	
	Factor		*1.6*	*1.5*	***5.0***	*1.5*	***1.9***	

Table 4.13: RMSE values in mm for check point residuals of direct (DG) and integrated georeferencing (IG) regarding two and four GCP groups, respectively.

4.2.4 Integrated Georeferencing without Ground Control Points

Determination of 3D GCP coordinates is a costly task, since usually performed by a surveying team on-site. Moreover, surveying work is often conducted in dangerous road environments. Hence, approaches that rely on no or just a minimal number of ground control points would be a great benefit. However, infrastructure applications typically demand 3D accuracies within the sub-decimeter range, and as shown in section 4.2.2, direct georeferencing solutions in urban environments usually degrade to several decimeters. Nonetheless, we demonstrate the feasibility of processing mobile mapping image sequences by only incorporating prior EOPs and calibrated ROPs into our modified COLMAP procedure.

Data Processing

We fixed all precalibrated ROPs among cameras and employed EOPs from direct georeferencing but no GCP observations, leading to mean residuals of projection centers from direct georeferencing close to 0. We conducted two processings for both mappings of the dataset Basel15, considering images captured at 448 epochs and 443 epochs, respectively. First, we incorporated images from the stereo cameras directed forward as well as from the backward facing panorama camera head. Second, we employed all twelve cameras, i.e. stereo forward imagery as well as five images per panorama camera. Regarding mean observations per image, there are significantly larger values for three cameras compared to twelve cameras

(see table 4.14). However, since mapped back and forth, basically all three cameras face sometime the same direction leading to more feature matches in the corresponding images.

	Seq. x.1 12 cameras	Seq. x.1 3 cameras	Seq. x.2 12 cameras	Seq. x.2 3 cameras
Registered images	5376	1344	5316	1329
3D points	1'730'085	590'716	1'730'226	573'834
Observations	12'181'922	4'354'391	11'538'702	3'981'022
Mean track length	7.0	7.4	6.7	6.9
Mean obs. per image	2266	3240	2171	2996
Mean reproj. error [px]	0.64	0.67	0.64	0.67

Table 4.14: COLMAP processing statistics for image sequences from forward stereo and multi-head panorama cameras of the dataset Basel15. While 12 cameras is the total amount, 3 cameras correspond to stereo forward and mono backward.

Check Point Investigations

We determined 3D coordinates of the check points used in table 4.13 by image measurements in single stereo pairs of the forward pointing cameras and by incorporating EOPs from both direct (DG) and integrated georeferencing (IG). Then we subtracted IG from DG values leading to our improvement values. Since twelve cameras merely delivered slightly better results, only values for investigations with three cameras are depicted in table 4.15. While mapping 1 shows a small horizontal improvement from 135 mm to 109 mm, the 2D RMSE value reduces significantly from 509 mm to 337 mm for mapping 2, due to a considerable accuracy improvement of the east component. The height RMSE value of mapping 1 lowers from 228 mm to 139 mm, mainly caused by sequence 2.1. However, induced by sequence 3.2, there is a small degradation from 463 mm to 495 mm for mapping 2. Even though 3D RMSE values for all but for 3.1 were improved, there are still mean values of 177 mm and 601 mm for mapping 1 and 2, respectively. However, values of sequences from the same mapping are more homogeneous compared to direct georeferencing. In summary, horizontal components can rather be improved than height values, especially if all height residuals point in the same direction and amount to several decimeters. Nevertheless, exploiting at least one reference height allows for mitigation of this issue.

[mm]	Seq.	ΔE	ΔN	ΔH	Δ2D	Δ3D	Seq.	ΔE	ΔN	ΔH	Δ2D	Δ3D
IG	1.1	89	50	116	102	155	1.2	80	331	563	340	658
	2.1	107	55	178	121	215	2.2	43	301	518	304	601
	3.1	86	59	121	104	160	3.2	45	364	403	367	545
	Mean x.1	94	55	**139**	**109**	**177**	Mean x.2	56	332	**495**	**337**	**601**
DG	1.1	74	59	138	95	168	1.2	465	24	605	466	764
	2.1	254	80	500	266	567	2.2	185	481	633	515	816
	3.1	27	34	46	44	64	3.2	196	511	150	548	568
	Mean x.1	118	58	**228**	**135**	266	Mean x.2	282	339	**463**	**509**	716
Impr.	1.1	-15	9	22	-7	13	1.2	385	-307	42	125	106
	2.1	147	24	322	145	351	2.2	142	180	115	211	215
	3.1	-59	-24	-75	-60	-96	3.2	152	147	-253	181	23
	Mean x.1	24	3	**89**	**26**	89	Mean x.2	226	7	**-32**	**172**	115
	Factor DG	*1.3*	*1.1*	*1.6*	*1.2*	*1.5*	*Factor DG*	*5.0*	*1.0*	*0.9*	*1.5*	*1.2*

Table 4.15: RMSE values in mm for check point residuals between integrated (IG) as well as direct georeferencing (DG) and tachymetry. Either 10 or 11 check points per sequence were utilized. Improvement (Impr.) values correspond to the difference between DG and IG. Only three cameras of the mobile mapping configuration were used, i.e. two cameras pointing forward and one directed backward.

Discussion

Integrated georeferencing with ROP self-calibration consistently revealed trajectory deviations from direct georeferencing in the order of several decimeters as well as multiple trajectory discontinuities. All image sequences captured in street sections 1 and 2 were processed using both two and four GCP groups. Additional support close to the respective curve led to an accuracy increase from 35 mm to 7 mm for the height component and from 62 mm to 33 mm in 3D. Compared to single forward stereo, utilization of all three stereo camera systems and two GCP groups for the dataset Basel14 improved the height accuracy from 18 mm to 8 mm and the 3D accuracy from 39 mm to 28 mm. Hence, our multi-stereo configuration showed the same performance as single stereo with additional GCP support in the middle of the curved road segments. Images captured in three horizontal directions at each epoch allow for strong tie point connections, which primarily stabilizes the height component.
We performed integrated georeferencing without GCPs for the dataset Basel15 employing both three and twelve cameras. Our minimal configuration consisting of a forward pointing stereo camera system and a backward facing camera delivered slightly poorer results. However, the need for only processing 25% of the total amount of images denotes an enormous efficiency increase. A resulting 3D accuracy of 177 mm for mapping 1 is acceptable, while 601 mm for mapping 2 does not meet our accuracy requirements. The main reason is a height offset of approx. 50 cm. Such constant deviations caused by direct georeferencing can only be eliminated by incorporation of at least one reference height.

4.3 Test Campaign in a Suburban Environment

Suburban environments with moderate GNSS coverage allow for direct georeferencing accuracies at the decimeter level. A degraded horizontal solution can be improved by integrated georeferencing without GCP utilization. However, uniformly deviated heights can only be corrected by incorporation of at least one reference point. We processed images of a road campaign in multiple scenarios, either exploiting imagery from one or opposite driving directions as well as including four or six cameras. Compared to an extended junction area in Basel, a larger study area featuring five loops was considered. Even though mainly focusing on integrated georeferencing without GCPs, we employed GCPs for ROP self-calibration.

4.3.1 Vehicle-Based Mobile Mapping System and Data

Mobile Mapping Platform

The vehicle-based stereovision mobile mapping system used for collecting the dataset Zug17 features six industrial cameras and a GNSS/IMU positioning system (see figure 4.20). It has a standard Y configuration consisting of forward pointing AVT stereo cameras as well as FHD stereo cameras facing back-right and back-left (see table 4.16). This is the same configuration as for the dataset Basel14, but a back-left pointing instead of a left pointing stereo camera system. Hence, this Y configuration allows for more potential connections between images captured in opposite driving directions. However, building façades parallel to the driving direction are not covered to the same degree in individual images. While the 11 MP cameras directed forward are separated by 1.054 m, the FHD cameras looking back-right and back-left have stereo bases of 0.747 m (see table 4.17). Same as for the dataset Basel14, a NovAtel SPAN inertial navigation system comprising a tactical grade inertial measurement unit and a L1/L2 GNSS kinematic antenna provides direct georeferencing.

Figure 4.20: Camera configuration that captured the dataset Zug17 (Source: iNovitas).

Camera type	Sensor size	Pixel size [µm]	Principal distance [mm]	Field of view [°]	Camera model
AVT	11 MP (4008 × 2672)	9.00	22.0	81 × 60	perspective
Basler/FHD	2 MP (1920 × 1080)	7.40	8.0	84 × 54	perspective

Table 4.16: Interior orientation parameters of the two different camera types used for processing the dataset Zug17.

Pointing direction of stereo cameras	forward	back-right	back-left
Camera type	AVT	Basler	Basler
Base length [mm]	1054	747	747

Table 4.17: Length of stereo bases for the standard Y camera configuration used for the campaign in Zug.

Study Area and Mobile Mapping Data

Our study area is located in the city center of Zug, Switzerland (see figure 4.21). While the eastern part of our study area shows mainly residential neighborhoods, there are several large commercial buildings close to the train station that create a challenging environment for GNSS positioning. We selected stereo images from five loops captured as part of a complete survey of the city of Zug on March 20 and 21, 2017. All loops were mapped in both directions and they build two test sites. Test site I is situated in the south-western part of our study area and features 1126 timestamps with a mean along-track spacing of 1.4 m. Test site II comprises 2531 epochs with a mean along-track distance between successive image exposures of 1.5 m.

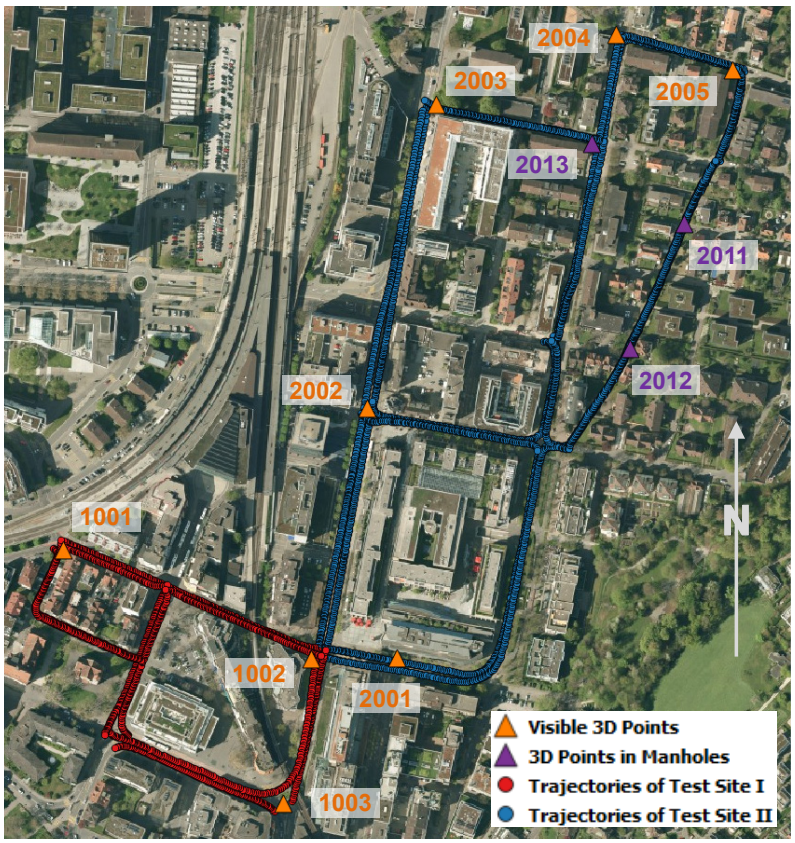

Figure 4.21: Orthophoto of our study area with overlaid trajectories and 3D reference points (Source of orthophoto: Geoportal Kanton Zug). Dimensions of depicted orthophoto section are ca. 630 m × 700 m.

Calibration and Direct Georeferencing

We performed calibration and direct georeferencing of the mobile mapping platform as described in section 3.3. Based on poses from direct georeferencing, we determined 3D coordinates for all reference points by employing four image observations for each computation. The resulting deviations between direct georeferencing and tachymetry are depicted in table 4.18 for test site I and in table 4.19 for test site II. Test site I shows a small 2D RMSE value of 29 mm and height deviations of ca. one decimeter. Test site II features a small RMSE value of 24 mm for the east component, but a larger value for the north component resulting in a 2D RMSE value of 91 mm. Height residuals are in the range of approx. one to two decimeters leading to a RMSE value of 142 mm. 3D RMSE values of 121 mm and 154 mm for test site I and test site II, respectively, mainly caused by deteriorated heights, are remarkably small considering the challenging suburban environment frequently causing GNSS degradations.

[mm]	ΔE	ΔN	ΔH	Δ2D	Δ3D
1001	-27	14	135	31	139
1002	34	-14	107	37	113
1003	10	-11	108	15	109
RMSE	26	13	**118**	**29**	**121**

Table 4.18: Check point residuals in mm between direct georeferencing and tachymetry as well as resulting RMSE values for test site I.

[mm]	ΔE	ΔN	ΔH	Δ2D	Δ3D
2001	-36	-8	88	37	96
2002	51	-98	-187	111	217
2003	-8	21	115	22	117
2004	-10	52	180	53	187
2005	-18	-5	111	18	112
2011	-8	-23	*170*	24	*171*
2012	-3	-218	*219*	218	*309*
2013	10	-15	*160*	18	*161*
RMSE	**24**	**88**	**142**	**91**	**154**

Table 4.19: Check point residuals in mm between direct georeferencing and tachymetry as well as resulting RMSE values for test site II. *Italic values* were excluded from RMSE computation due to 3D points in manholes with indeterminable accurate heights.

Reference Data

We used several highly accurate cadastral reference points for our investigations. Eight of them are signalized using bolts and thus clearly defined in the mobile mapping imagery (see orange 3D points in figure 4.21 and figure 4.22). In contrast, the other three reference points are protected in manholes ca. 10-20 cm below the road surface (see purple 3D points in figure 4.21 and figure 4.22). Since we determined 3D coordinates of the cover plate centers using stereo images, they deviate from the provided 3D reference coordinates in the centimeter range for 2D and in the decimeter range for the height component. All measurements were performed in the Swiss horizontal reference frame LV03 and in the LN02 vertical reference frame based on leveled heights.

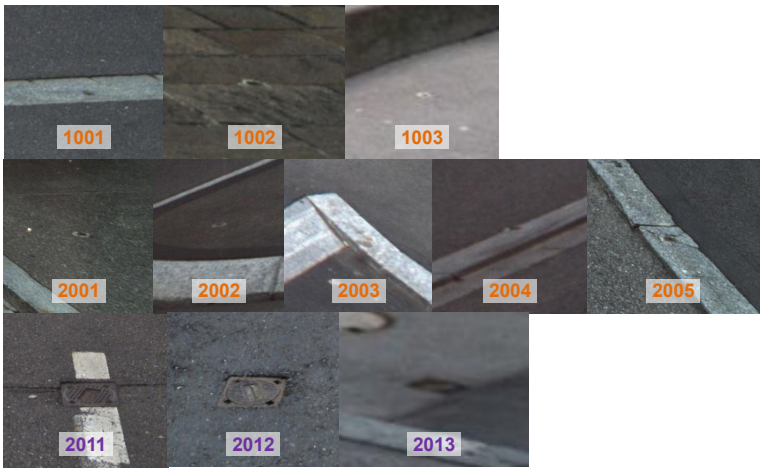

Figure 4.22: Image sections of all 11 3D reference points used for the dataset Zug17. Top row: 3D points in test site I, middle row: visible 3D points in test site II, bottom row: 3D points in manholes in test site II.

4.3.2 Integrated Georeferencing with Ground Control Points and ROP Self-Calibration

Data Processing

In order to refine all precalibrated ROPs among cameras, we employed image observations to three GCPs for each test site. Several thousand images per test site were processed (see table 4.20), and in both cases images from opposite driving directions. However, compared to six cameras for test site I, we only considered four cameras for test site II, i.e. stereo forward and mono images from both back-right and back-left (left cameras of stereo systems). This is the main reason for smaller values of mean track length and mean observations per image (6.4 vs. 7.4 and 1676 vs. 2306, respectively).

	Test site I	Test site II
Epochs	1126	2531
Cameras	6	4
Registered images	6751	10'085
3D points	2'107'485	2'622'882
Observations	15'564'304	16'900'128
Mean track length	7.4	6.4
Mean obs. per image	2306	1676
Mean reproj. error [px]	0.60	0.65

Table 4.20: COLMAP processing statistics for ROP self-calibration in Zug.

Check Point Investigations

We computed 3D point coordinates using measurements carried out in two stereo image pairs, and calculated deviations to reference coordinates from tachymetry. There is a RMSE value of 20 mm for both the east and north component leading to a 2D RMSE value of 28 mm for test site I (see table 4.21). The corresponding height RMSE value is half as large, however, all RMSE values for GCP residuals of test site I are significantly larger compared to test site II (see table 4.22). This is probably due to a smaller extent and thus a stronger network geometry for test site I. We selected five check points for test site II, but only two of them (points 2002 and 2004) have clearly defined heights and were thus considered for height RMSE computation. The resulting 2D RMSE value amounts to 33 mm, and the height RMSE value is approx. one decimeter.

[mm]	ΔE	ΔN	ΔH	Δ2D	Δ3D
1001	5	6	-2	8	8
1002	25	32	12	41	43
1003	24	-9	10	25	27
RMSE	20	20	9	28	**30**

Table 4.21: Ground control point residuals in mm between integrated georeferencing and tachymetry as well as resulting RMSE values for test site I and ROP self-calibration I.

[mm]	ΔE	ΔN	ΔH	Δ2D	Δ3D
2001	3	-4	3	5	6
2002	-11	12	-52	16	54
2003	2	-7	7	7	10
2004	-8	7	129	11	130
2005	3	-1	-1	3	3
2011	-27	-13	*132*	30	*136*
2012	20	-5	*163*	20	*164*
2013	27	-55	*142*	62	*155*
RMSE all	16	21	62	26	63
RMSE **GCPs**	3	5	5	5	**7**
RMSE CPs	20	26	**98**	**33**	**99**

Table 4.22: Both ground control point (bold numbers) and check point residuals in mm between integrated georeferencing and tachymetry as well as resulting RMSE values for test site II and ROP self-calibration II. *Italic values* were excluded from RMSE computation due to 3D points in manholes with indeterminable accurate heights.

4.3.3 Integrated Georeferencing without Ground Control Points

Data Processing

We fixed all ROPs among cameras and exploited EOPs from direct georeferencing but no GCP observations. We performed several processings for both test site I and test site II. In case of test I, we considered six but also four cameras, one and opposite driving directions, and ROPs from self-calibration I. For test site II, we always used four cameras but one and opposite driving directions. Furthermore, we exploited ROPs from self-calibration I as well as from self-calibration II.

Check Point Investigations

We determined 3D check point coordinates by measuring each point in four images, and computed deviations to reference coordinates from tachymetry. Table 4.23 compares the following three scenarios: images from six cameras and opposite driving directions (scenario I, top), images from six cameras and one driving direction (scenario II, middle) as well as images from four cameras and one driving direction (scenario III, bottom). Scenario I delivers the best RMSE values, and there are only slight RMSE differences between scenarios II und III. However, all 2D RMSE values are larger than the surprisingly very accurate value of 29 mm for direct georeferencing. In order to eliminate an apparent height offset in direct georeferencing, we translated the computed integrated georeferencing solution to the 3D reference coordinates of control point 1002. This led to a small height RMSE value of 15 mm for scenario I, mainly caused by homogeneous height deviations that is represented with a standard deviation value of 11 mm. Height RMSE values for scenarios II and III are approximately twice as large, and 3D RMSE values for all scenarios range from 35 mm to 47 mm. In summary, incorporation of imagery from opposite driving directions is more beneficial than employing stereo instead of mono images from camera systems directed back-right and back-left. Moreover, using at least one reference point is essential in order to obtain height accuracies at the centimeter level and thus 3D accuracies of better than one decimeter.

[mm]		ΔE	ΔN	ΔH	Δ2D	Δ3D
No GCPs, IG-TPS, 6 cameras, back & forth, 6751 images (scenario I)	1001	-2	13	102	13	103
	1002	-2	-4	85	4	85
	1003	-32	-45	106	56	119
	Mean	-12	-12	97	24	102
	SD	18	30	**11**	28	17
	RMSE	19	27	**98**	**33**	**103**
	Factor DG	*1.4*	*0.5*	*1.2*	*0.9*	*1.2*
	IG 1002	18	26	**15**	31	**35**
	Factor DG	*1.4*	*0.5*	***7.9***	*0.9*	***3.5***
No GCPs, IG-TPS, 6 cameras, **forth**, 3662 images (scenario II)	1001	-7	-3	114	8	114
	1002	-18	-7	81	19	83
	1003	-50	-50	116	71	136
	Mean	-25	-20	104	33	111
	SD	22	26	**20**	34	27
	RMSE	31	29	**105**	**43**	**113**
	Factor DG	*0.8*	*0.4*	*1.1*	*0.7*	*1.1*
	IG 1002	19	25	**28**	32	**42**
	Factor DG	*1.4*	*0.5*	***4.2***	*0.9*	***2.9***
No GCPs, IG-TPS, **4 cameras**, **forth**, 2438 images (scenario III)	1001	-13	6	117	15	118
	1002	-5	-8	78	9	79
	1003	-42	-47	123	63	138
	Mean	-20	-16	106	29	111
	SD	19	28	**24**	30	30
	RMSE	26	28	**108**	**38**	**114**
	Factor DG	*1.0*	*0.5*	*1.1*	*0.8*	*1.1*
	IG 1002	22	24	**34**	33	**47**
	Factor DG	*1.2*	*0.5*	***3.5***	*0.9*	***2.6***
Direct georeferencing (DG)	RMSE	26	13	118	29	121

Table 4.23: Check point residuals in mm between integrated georeferencing (IG) exploiting ROPs from self-calibration I and tachymetry (TPS) for test site I. Besides, resulting mean, standard deviation (SD) and RMSE values as well as improvement factors compared to direct georeferencing (DG) are shown. IG 1002 indicates RMSE values if the integrated georeferencing solution is translated to 3D reference coordinates of point 1002.

Again for test site II, we computed 3D check point coordinates using four observations per point, and calculated deviations to reference coordinates from tachymetry. As shown by table 4.24 and table 4.25, there are larger height RMSE values for back and forth compared to direct georeferencing, while incorporation of images from only one direction leads to slightly better height accuracies. However, 2D RMSE values reduce in case of all scenarios. In order to remove an obvious systematic height effect in direct georeferencing as well as to show the feasibility of obtaining 3D deviations within the sub-decimeter range, we performed a 3D translation to control point 2003. This resulted in a horizontal accuracy decrease for self-calibration I, but led to smaller 2D RMSE values for self-calibration II. Height accuracies significantly increased for opposite driving directions (from 158 mm to 25 mm and from 156 mm to 21 mm), while improving by a factor of ca. 2 if only employing images captured in one direction. The main reason for this difference are more homogeneous height residuals and thus smaller standard deviation values for two driving directions. As expected and shown by the left part of table 4.25, incorporation of imagery from back and forth as well as exploiting ROPs calibrated on the same test site feature the most stable inner geometry, which leads to the best accuracies. To sum up, horizontal accuracies of less than one decimeter are possible without reference data, however, constant height offsets from direct georeferencing can only be corrected by incorporation of at least one reference point.

[mm]	ΔE	ΔN	ΔH	Δ2D	Δ3D	[mm]	ΔE	ΔN	ΔH	Δ2D	Δ3D
2001	-44	-92	152	102	183	2001	-20	-89	65	91	112
2002	-40	-22	113	46	122	2002	-19	-14	70	24	74
2003	10	44	157	45	163	2003	40	37	161	54	170
2004	-3	104	189	104	216	2004	3	98	232	98	252
2005	21	33	168	39	172	2005	30	41	85	51	99
2011	-36	-81	*130*	89	*157*	2011	-21	-59	*111*	63	*127*
2012	-7	-56	*205*	56	*213*	2012	9	-52	*161*	53	*170*
2013	40	-10	*207*	41	*211*	2013	57	-18	*202*	59	*210*
Mean	-7	-10	156	65	171	Mean	10	-7	123	62	141
SD	31	67	**28**	28	34	SD	30	62	**72**	24	71
RMSE	**30**	**64**	**158**	**70**	**174**	RMSE	**30**	**58**	**139**	**65**	**155**
DG	24	88	142	91	**154**	DG	24	88	142	91	**154**
Factor DG	*0.8*	*1.4*	*0.9*	***1.3***	*0.9*	*Factor DG*	*0.8*	*1.5*	*1.0*	***1.4***	*1.0*
IG 2003	**33**	**83**	**25**	**90**	**84**	IG 2003	**41**	**72**	**75**	**83**	**109**
Factor DG	*0.7*	*1.1*	***5.7***	*1.0*	*1.8*	*Factor DG*	*0.6*	*1.2*	***1.9***	*1.1*	*1.4*
4 cameras, back & forth, 10'085 images						4 cameras, forth, 5004 images					

Table 4.24: Check point residuals in mm between integrated georeferencing exploiting ROPs from **self-calibration I** and tachymetry for test site II. Besides, resulting mean, standard deviation (SD) and RMSE values as well as improvement factors compared to direct georeferencing (DG) are shown. IG 2003 indicates RMSE values if the integrated georeferencing solution is translated to 3D reference coordinates of point 2003. *Italic values* were excluded from RMSE computation due to 3D points in manholes with indeterminable accurate heights.

[mm]	ΔE	ΔN	ΔH	Δ2D	Δ3D	[mm]	ΔE	ΔN	ΔH	Δ2D	Δ3D
2001	-29	-15	148	33	151	2001	-9	-16	60	18	63
2002	-33	1	120	33	124	2002	-7	0	83	7	83
2003	-18	-32	155	36	159	2003	9	-39	164	40	168
2004	-25	-30	181	39	185	2004	-18	-33	200	38	203
2005	-15	-57	168	60	178	2005	-2	-52	86	52	101
2011	-11	-43	*141*	44	*147*	2011	-1	-47	*121*	47	*130*
2012	10	-35	*202*	37	*205*	2012	24	-37	*160*	44	*166*
2013	13	-78	*205*	79	*220*	2013	29	-89	*190*	93	*211*
Mean	-14	-36	154	45	160	Mean	3	-39	119	42	124
SD	17	24	**23**	16	24	SD	16	26	**60**	25	60
RMSE	**21**	**43**	**156**	**47**	**161**	RMSE	**16**	**46**	**130**	**49**	**135**
DG	24	88	142	91	**154**	DG	24	88	142	91	**154**
Factor DG	*1.1*	*2.0*	*0.9*	***1.9***	*1.0*	*Factor DG*	*1.5*	*1.9*	*1.1*	***1.9***	*1.1*
IG 2003	**17**	**23**	**21**	**28**	**30**	IG 2003	**16**	**24**	**70**	**29**	**75**
Factor DG	*1.4*	*3.8*	***6.8***	***3.3***	***5.1***	*Factor DG*	*1.5*	*3.7*	***2.0***	***3.1***	***2.1***
4 cameras, back & forth, 10'085 images						4 cameras, forth, 5004 images					

Table 4.25: Check point residuals in mm between integrated georeferencing exploiting ROPs from **self-calibration II** and tachymetry for test site II. Besides, resulting mean, standard deviation (SD) and RMSE values as well as improvement factors compared to direct georeferencing (DG) are shown. IG 2003 indicates RMSE values if the integrated georeferencing solution is translated to 3D reference coordinates of point 2003. *Italic values* were excluded from RMSE computation due to 3D points in manholes with indeterminable accurate heights.

Discussion

Integrated georeferencing without GCPs resulted in 2D accuracies of 33-43 mm for test site I and 47-70 mm for test site II. Compared to direct georeferencing, this is a slight decrease for test site I and an improvement by a factor of 1.3-1.9 for test site II. However, a 2D accuracy of 29 mm for test site I is extraordinary. Heights did not change significantly by integrated georeferencing exploiting EOPs and ROPs. The main reason are similar height residuals pointing in the same direction. Incorporation of one control point and thus removing a constant height offset led to 3D accuracies of better than 5 cm for test site I and 30-109 mm for test site II. Employing mono back-right and mono back-left instead of stereo images does hardly lead to an accuracy decrease. However, processing imagery captured in opposite driving directions results in more homogeneous heights. To conclude, 3D accuracies within the sub-decimeter range are feasible by incorporation of only one control point. In case of high direct georeferencing accuracies for the horizontal components as encountered in our experiments, even a height reference suffices. Nevertheless, at least one additional check point is recommended in order to guarantee a reliable solution.

4.4 Test Campaign at a Train Station

Compared to urban road environments, GNSS availability on railway tracks is usually sufficient, which allows for 3D accuracies from direct georeferencing within the sub-decimeter range. However, feature matching in such rail environments poses a big challenge. If considering images captured by a forward pointing camera, the top half mainly contains sky areas that are not of interest. Since omnipresent and distinctive, most of the features will be extracted in track ballast regions, which results in many tie points constituting approximately a plane in 3D space. In order to prevent such weak connections, we selected a train station showing several distinctive structures. This enables a uniform feature point distribution per image, but also adequate tie point connections in case of large viewpoint changes. Hence, we demonstrate the feasibility and accuracy potential of self-calibrating ROPs among individual stereo systems by only employing calibrated stereo bases and prior camera poses from direct georeferencing.

4.4.1 Train-Based Mobile Mapping System and Data

Mobile Mapping Platform

The multi-sensor stereovision mobile mapping system assembled for capturing the dataset Vienna16 features ten industrial cameras, a GNSS/IMU positioning system and a laser scanner (see figure 4.23). All sensors are mounted on aluminum beams attached to a locomotive, not on the roof but lower in front of the locomotive shell, allowing to map the complete rail corridor. Two stereovision systems are directed forward, one comprising 11 MP RGB cameras and the others 4 MP grayscale cameras (see table 4.26 and table 4.27). The stereo system with grayscale cameras has a smaller base of 0.925 m compared to 1.268 m for forward RGB, and is mainly used for tunnels, which are additionally illuminated by two floodlights. The other three stereo systems comprise FHD cameras, one facing downward to the rails for efficient inspection and the others pointing right or left, which is similar to Basel14. Same as the MMS that captured the datasets Basel14 and Basel15 as well as Zug17, a NovAtel SPAN inertial navigation system consisting of a tactical grade inertial measurement unit and a L1/L2 GNSS kinematic antenna provides direct georeferencing. A high-end laser scanner directed downward enables analysis of rails, e.g. rail wear.

Figure 4.23: Mobile mapping system assembled for a rail campaign in Vienna (Source: iNovitas).

Camera type	Sensor size	Pixel size [µm]	Principal distance [mm]	Field of view [°]	Camera model
AVT	11 MP (4008 × 2672)	9.00	21.1	81 × 60	perspective
Grayscale	4 MP (2048 × 2048)	5.50	7.8	72 × 72	perspective
Basler/FHD	2 MP (1920 × 1080)	7.40	7.8	84 × 54	perspective

Table 4.26: Interior orientation parameters of the three different camera types used for processing the dataset Vienna16.

Pointing direction of stereo cameras	forward RGB	forward gray	down	right	left
Camera type	AVT	Grayscale	Basler	Basler	Basler
Base length [mm]	1268	925	820	752	757

Table 4.27: Length of stereo bases for the camera configuration assembled for a rail campaign in Vienna.

Study Area and Mobile Mapping Data

Our test site is located in Vienna, Austria, and comprises the train station called Wien Erzherzog-Karl-Strasse, situated north-east of the larger station Wien Stadlau. Multi-stereo imagery was captured twice on the same track but from opposite driving directions on October 11, 2016. We considered a track section length of approx. 197 m leading to 98 (15:19:21-15:19:41, 20 seconds) and 87 (16:03:38-16:03:56, 18 seconds) timestamps, respectively (see figure 4.24). Although rather short, the employed imagery contains several distinctive structures that is often not the case in arbitrary rail mapping images. Since captured from ten cameras, these 185 epochs resulted in 1850 images. Mean along-track distances between successive image exposures are 2.0 m and 2.3 m, respectively. While there is a departure platform with facilities and roofing covering the south-western part on the one side of the utilized track, the other side features more free space due to another track as well as vegetation and some buildings lying farther away (see figure 4.25).

Calibration and direct georeferencing were basically performed according to section 3.3, but in the coordinate reference system MGI / Austria GK East. Interior orientation parameters as well as offsets and rotations of individual stereo bases were precisely calibrated in an indoor calibration field with many signalized coded targets. However, ROPs among stereo systems as well as lever arm and misalignment were calibrated on-site with medium precision.

Figure 4.24: Orthophoto of the study area with overlaid blue trajectories and orange 3D check points (left) (Source of orthophoto: Stadt Wien - data.wien.gv.at). Furthermore, there are mobile mapping images from the left camera of the forward pointing stereo system with overlaid orange 3D check points (right).

Figure 4.25: Images of left cameras of all stereo systems captured at the same location.

4.4.2 Integrated Georeferencing with ROP Self-Calibration using Fixed Stereo Bases

Data Processing

There were no GCP observations, so that we basically incorporated EOPs from direct georeferencing. Moreover, we fixed stereo bases of individual camera systems, but COLMAP self-calibrated ROPs between individual stereo systems. Hence, precalibrated stereo bases defined the correct object scene scale. Nonetheless, accurate initial EOPs from direct georeferencing were also crucial for ROP estimation among individual bases.

COLMAP processing resulted in a mean value of 2240 observations per image and a mean track value of 5.4 (see table 4.28), even though images from the stereo system directed downward contain significantly fewer features. For another rail dataset (Yconfig16) featuring a standard Y configuration, which corresponds to Zug17 and comprises forward looking AVT stereo cameras as well as FHD stereo cameras directed back-right and back-left, COLMAP computed an even smaller value of 1655 for mean observations per image. The principal reasons are rather large mean along-track spacings of 3.4 m (64 epochs) and 3.8 m (56 epochs), respectively.

	Vienna16	Yconfig16
Registered images	1850	720
3D points	770'752	234'086
Observations	4'143'899	1'191'755
Mean track length	5.4	5.1
Mean obs. per image	2240	1655
Mean reproj. error [px]	0.72	0.61

Table 4.28: COLMAP processing statistics for ROP self-calibration in case of two rail datasets.

Analysis of Connectivity Matrix

Figure 4.26 depicts 3D tie point connections after bundle adjustment, not only established between images of the same stereo camera system but also between imagery captured by different stereovision systems. Rows and columns represent all processed images in ascending order, from top to bottom and from left to right. Regarding the succession of forward RGB that is shown in the dashed gray square in the top left corner, there are first images captured by the left camera of the forward pointing system in direction 1, then in opposite direction 2, followed by imagery from the right camera in direction 1 and eventually right images in direction 2. Numbers of feature matches are color coded and a threshold of 30 defines the transition from red to blue.

Images recorded from forward facing cameras in the same driving direction have a great many connections, i.e. forward left and right but also forward RGB and gray. However, there are barely any connections between forward imagery captured from opposite driving directions. Some matches were obtained between images from the forward looking cameras and images from the cameras directed downward, right and left. Considering images from the downward pointing cameras, there are only a few connections between consecutive images of the same camera as well as between left and right camera images at the same epoch. The main reason are short distances to mapping objects such as rails, crossties and track ballast resulting in small acquired areas and small image overlapping, i.e. same points are seen in at most two consecutive images. Due to complementary mapped regions, there are no connections between downward facing imagery and images captured by the right as well as the left stereovision systems. Images from the right stereo camera system are well connected with images from the left stereo system captured in the opposite direction.

Besides, figure 4.26 enables a comparison between the assembled stereo camera configuration and a standard Y camera configuration often used for road mapping, e.g. dataset Zug17. The Vienna16 configuration allows for only a few tie points between the forward facing stereovision systems and the stereo camera systems directed right as well as left (see green rectangles). In contrast, imagery from the forward pointing stereovision system is well connected with images captured in the opposite direction from the systems looking back-right and back-left. The prime reason are stereo system pointing differences of 90° for Vienna16 and ca. 45° for the standard Y configuration. Still remarkable that several matches were established, since SIFT features struggle with viewing direction differences of more than 30° (Hartmann et al., 2016). On the other hand, many more SIFT features can be matched between stereovision systems pointing right and left compared to stereovision systems facing back-right and back-left (see purple squares). While left and right images from opposite driving directions have approximately the same viewing direction, there is a difference of around 90° between back-right and back-left stereo camera systems.

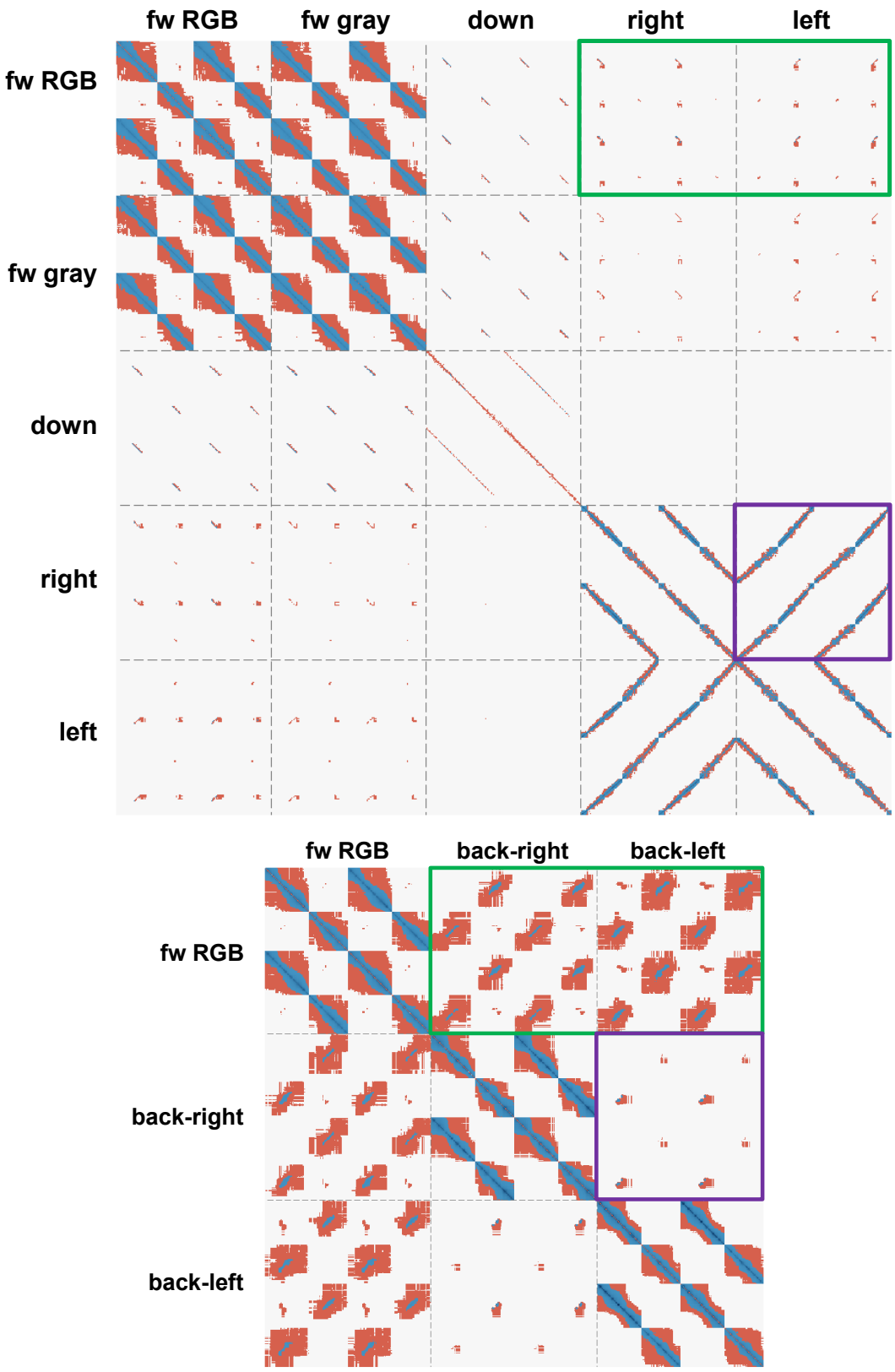

Figure 4.26: Connectivity matrices for Vienna16 (top) and for another rail dataset with a standard Y camera configuration (bottom). Colors represent tie point matches: from red that is up to 30 connections over to light blue until dark blue, which stands for several hundreds up to thousands of connections.

Check Point Investigations

Since no reference points were available for the dataset Vienna16, we selected three natural points defined by markings on the departure platform (see figure 4.24). We determined 3D coordinates of these check points based on multiple stereo image pairs captured consecutively, i.e. four observations per 3D point calculation. First, we computed our three references by averaging 3D point coordinates determined from the two opposite driving directions based on forward RGB. Second, we estimated six 3D point coordinates based on multiple stereo images from forward gray, i.e. back and forth for all three points. Moreover, we determined 3D point coordinates employing imagery from the stereovision systems facing right and left, respectively, only captured in the corresponding direction. Eventually, we computed RMSE values for check point residuals between 3D coordinates determined by forward RGB and the other three stereo camera systems.

We obtained exactly the same results for the right and left stereovision systems (see table 4.29). Values of up to 10 mm per component lead to a 3D RMSE value of 15 mm. While the height value is even better for forward gray compared to right and left, the degraded horizontal component values cause a 3D RMSE value of 23 mm. This even though forward RGB and forward gray have the same pointing direction and hence many connections. Reasons might be a lower geometric resolution compared to forward RGB and some image blur in the grayscale images. Nonetheless, the feasibility of self-calibrating ROPs among individual stereo systems leading to a 3D accuracy at the centimeter level has been shown. Therefore, outdoor calibration fields for ROP estimation are optional, but at least some GCPs for boresight alignment are mandatory.

[mm]	# CPs	ΔX	ΔY	ΔH	Δ2D	Δ3D
forward RGB-right	3	10	8	7	13	15
forward RGB-left	3	10	8	7	13	15
forward RGB-forward gray	6	14	17	5	22	23

Table 4.29: RMSE values in mm for check point residuals between 3D coordinates determined by forward RGB and other stereo camera systems.

Discussion

Integrated georeferencing with ROP self-calibration utilizing GCPs is a standard procedure. However, we did not rely on any GCPs, but on precisely calibrated stereo bases in order to obtain metric information. Self-calibrated ROPs among the stereo camera systems pointing forward as well as right and left enabled relative 3D point accuracies in object space of 15 mm. While satisfactory for many applications, this procedure is efficient since not requiring any calibration field on-site. We further compared two camera configurations in terms of tie point matching. Images from a forward pointing camera can be well connected with images captured by a camera directed back-right or back-left, but in the opposite driving direction. However, only a moderate number of matches can be established between back-right and back-left. In contrast, images from right and left pointing cameras acquired in opposite driving directions can build strong tie point connections, while connecting forward with right or left is more challenging.

4.5 Test Campaigns in a Building

Indoor environments are even more challenging than outdoor scenes. As there is no GNSS coverage in buildings, initial camera poses are computed by lidar SLAM. Furthermore, feature extraction and matching is aggravated due to weakly textured surfaces and repetitive patterns. We mapped the same floor of a building using our portable MMS twice. While one horizontal laser scanner was available for the first campaign, an additional laser scanner was employed for the second indoor mapping. Compared to outdoors, no stereo imagery was captured and images of the multiple panorama camera heads do barely overlap. We show that our integrated georeferencing approach based on COLMAP is also able to successfully process challenging datasets collected in an indoor environment. Besides, we determine the accuracy potential of both lidar SLAM and integrated georeferencing.

4.5.1 Indoor Mobile Mapping System and Data

Portable Mobile Mapping System

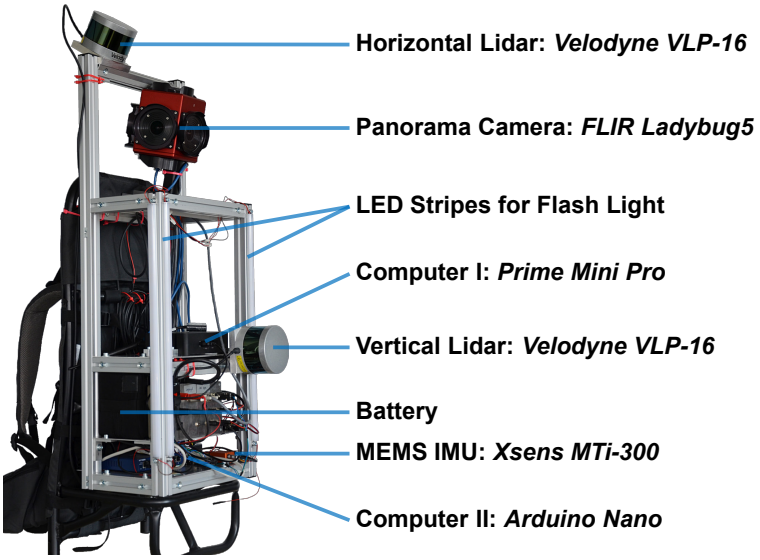

Figure 4.27: Sensor configuration of the portable panoramic mobile mapping system.

The portable panoramic mobile mapping system of the Institute of Geomatics at FHNW consists of multiple sensors mounted on a rigid aluminum frame, which is attached to a backpack (see figure 4.27). A multi-head 360° panorama camera of the type FLIR Ladybug5 serves for image capturing and was calibrated according to section 3.3.1. Each of the six camera heads of the Ladybug5 camera has a resolution of 2448 x 2048 pixels (5 MP) at a pixel size of 3.45 µm, a principal distance of 4.3 mm and a FOV of about 113° x 94°, hence featuring ultra-wide-angle optics. The labeling of the five horizontally arranged camera heads is depicted in figure 4.28, starting with cam1 facing backward and increasing in clockwise direction. The panorama camera is tilted by a few degrees in figure 4.27, but the arrangement of the five consecutive camera heads turns almost horizontal in the case the backpack is carried by a person. The MMS uses two multi-profile Velodyne VLP-16 laser scanners (lidar PUCK) with a 360° horizontal FOV and a 30° vertical FOV for navigation as well as for mapping (see table 4.30). The horizontal lidar is mounted on top of the frame. It is tilted by approx. 30 degrees in order to not only map walls but also some points on floors as well as on ceilings. Even more 3D points of horizontal surfaces are captured by the vertical lidar, which additionally improves the performance of lidar SLAM. The MEMS based Xsens IMU of the type MTi-300 further supports 3D lidar SLAM. For dynamic use, the accuracy of the attitude angle roll γ is specified with 0.3°, pitch θ and heading ψ with 1.0°. In addition, there is an on-board computer for data processing and storage (Prime Mini Pro), a computer for synchronization (Arduino

Nano), a battery for power supply as well as four LED stripes for illumination on the backpack. Our acquisition software was implemented using the Robot Operating System (ROS) framework. Detailed information about our portable mobile mapping system is given in Blaser et al. (2018).

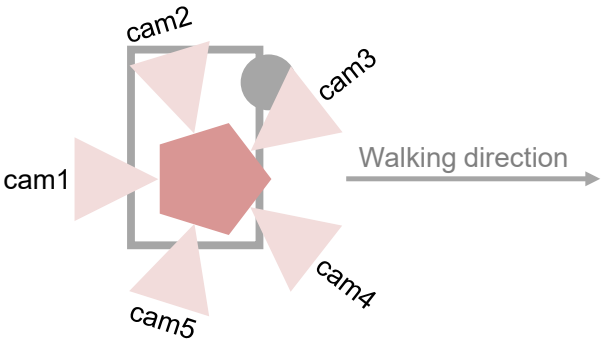

Figure 4.28: Top view of the backpack with camera head naming convention.

Field of view	360° × 30° (±15°)
Channels	16
Points per second	ca. 300'000
Range	up to 100 m
Accuracy	±3 cm

Table 4.30: Specifications of laser scanners mounted on the portable mobile mapping system.

Test Site and Data

Our indoor study area is located on the sixth floor of the former main campus building of the University of Applied Sciences and Arts Northwestern Switzerland in Muttenz nearby Basel. As depicted in figure 4.29 and figure 4.32, it features a hallway that has a dimension of ca. 27 m x 24 m, and leads to several offices, lecture rooms, laboratories, two staircases as well as five elevators. The typical corridor width amounts to approx. 3 m, with a minimum of about 2 m (left part, no. 6 in figure 4.29) and a maximum of 4 m (right part, no. 3 in figure 4.29).

We performed two campaigns, the first in November 2017 (27.11.2017 17:17-17:37, see figure 4.29) and the second in March 2018 (21.03.2018 10:50-11:14, see figure 4.32). In both cases, we started data acquisition at the origin of the local geodetic coordinate system marked with a dark gray diamond and we set the camera trigger constraints to a distance interval of 1 m and an azimuth change of 15 degrees. Figure 4.30 shows the mapping area at a specific camera position, and gives an impression of the difficult lighting conditions as well as poor texture. Please note that our portable mobile mapping system allows for complete indoor initialization. While we incorporated two laser scanners for the campaign in March 2018, we only used the horizontal laser scanner for the campaign in November 2017. During the complete mapping process, online 3D lidar SLAM using the Google Cartographer (Hess et al., 2016) based on laser scanner and IMU data was performed. This real-time computation of camera poses and 3D points served for user guidance as well as for geometrically constrained camera triggering.

We determined 3D coordinates of many reference points representing natural markings e.g. on door frames, elevator and room corners (see figure 4.31) by tachymetry. These reference points have an accuracy of approximately 5 mm and we used several of them for our indoor investigations as ground control points or check points.

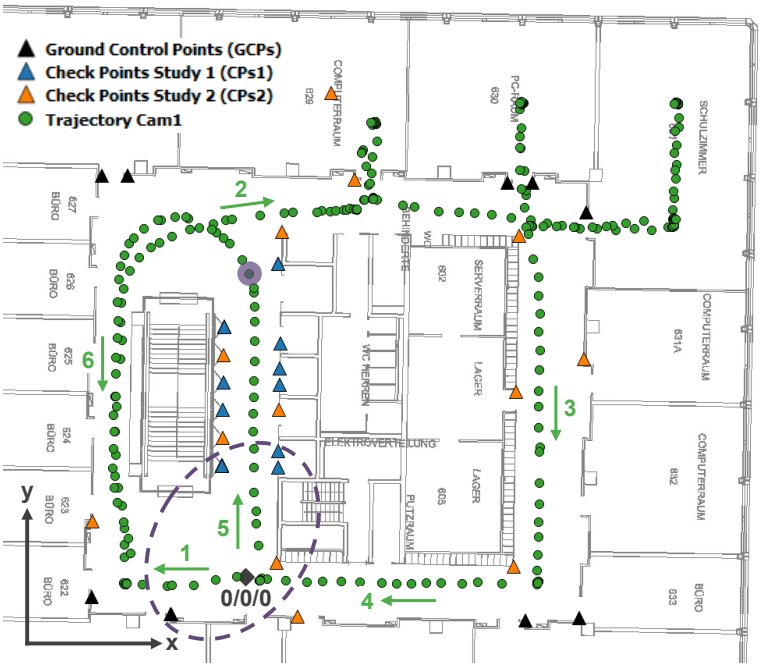

Figure 4.29: Floor base map with overlaid projection centers of camera head cam1 for the campaign performed in November 2017, ground control points, check points, local geodetic coordinate system, area covered by figure 4.31 (purple dashed ellipse) and MMS location at the time of capturing the images of figure 4.30 (purple filled circle).

Figure 4.30: Images captured at the same location (see figure 4.29) from the upward facing camera head cam6 (bottom right), backward pointing camera head cam1 (top left) and the consecutive camera heads cam2 (top middle), cam3 (top right), cam4 (bottom left), cam5 (bottom middle).

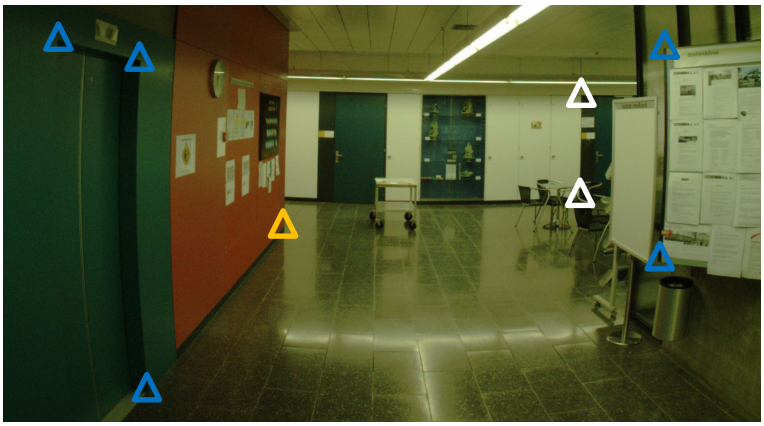

Figure 4.31: Image section from camera head cam1 with marked ground control points (white) as well as check points (blue and orange).

Figure 4.32: Floor base map with overlaid projection centers of camera head cam1 for the campaign performed in March 2018, ground control points, check points and local geodetic coordinate system (2GCPs/2CPs: two points at the same location, but on different height levels, e.g. 3D points on door frames).

4.5.2 Integrated Georeferencing with Ground Control Points

Data Processing

For the campaign performed in November 2017, we selected three GCPs in every corner and measured sensor coordinates for each of these 12 GCPs in three consecutive images of camera head cam2. Where only two GCPs per group are visible in figure 4.29, two of them lie at the same 2D position but on two different height levels (see figure 4.31). We fixed the precalibrated ROPs for our indoor processings, exploited previously computed EOPs from lidar SLAM, and COLMAP extracted DSP-SIFT features. Since the images from the upward facing camera head cam6 predominantly contain homogeneous surfaces leading to few feature correspondences, we only processed images from the horizontal pointing camera

heads cam1-cam5 captured at 270 locations (see figure 4.33) for the dataset Muttenz17 and at 223 locations for Muttenz18. As depicted in table 4.31, this resulted in 1350 registered images for Muttenz17 and 1115 registered images for Muttenz18. We obtained long mean 3D point tracks of 9.8 and 8.4, respectively. However, mean observations per image of less than 700 are moderate compared to typically at least three times larger values for outdoor environments.

We performed all experiments on a Linux laptop with an Intel Xeon E3-1535M 8-Core processor (2.9 GHz), 32 GB RAM and a Nvidia Quadro M2000M graphics card. For the dataset Muttenz17, this laptop required 40.2 minutes for feature extraction, 131.5 minutes for feature matching and 29.6 minutes for bundle adjustment. The considerably longer feature extraction time (feature extraction to matching ratio of ca. 1:3) compared to our outdoor datasets (feature extraction to matching ratio of ca. 1:10) is due to DSP-SIFT instead of SIFT, yet leading to a larger number of features. However, these features and the resulting feature correspondences are not necessarily more evenly distributed all over the images.

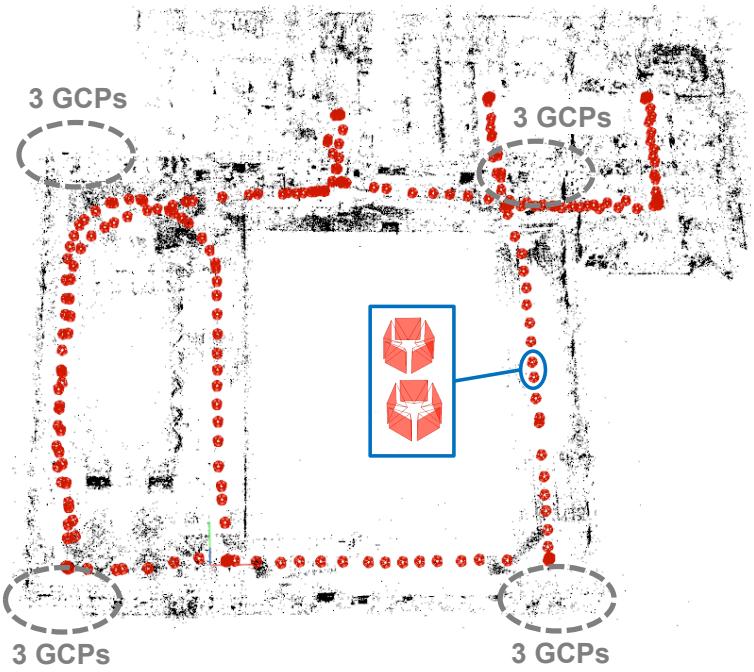

Figure 4.33: Georeferenced mobile mapping images (red) and 3D tie points (black) at our indoor test site using our modified COLMAP processing pipeline for Muttenz17.

	Muttenz17	Muttenz18
Registered images	1350	1115
3D points	88'955	92'597
Observations	876'096	779'782
Mean track length	9.8	8.4
Mean obs. per image	649	699
Mean reproj. error [px]	0.80	0.79

Table 4.31: COLMAP processing statistics of indoor image sequences.

Check Point Investigations for Building Dataset Muttenz17

We performed two studies, both evaluating the absolute 3D measurement accuracy of our portable mobile mapping system. Study 1 aimed at assessing the calibrated relative orientation parameters among all camera heads. Hence, we selected six blue check points close to the elevators at three locations on two

different height levels for each camera head (see figure 4.29 and figure 4.31). Then, we determined 3D check point coordinates by image measurements in four consecutive images and computed residuals to tachymetry per camera head (see table 4.32). The RMSE values for 3D check point residuals vary from 15 to 20 mm with a mean value of 17 mm. Thus, measurements for 3D point determination can be performed in arbitrary camera heads without accuracy degradation that is of high practical relevance.

Camera head	# CPs	Δx [mm]	Δy [mm]	Δz [mm]	Δ3D [mm]
cam1	6	10	11	8	17
cam2	6	12	7	9	17
cam3	6	10	6	8	15
cam4	6	8	9	12	16
cam5	6	10	13	11	20

Table 4.32: RMSE values of study 1 for check point residuals between integrated georeferencing and tachymetry.

For study 2, we determined the 3D coordinates of the 13 orange check points distributed all over the hallway by image measurements in four consecutive images of camera head cam3 or cam4 (see figure 4.29). The resulting RMSE value for 3D check point residuals between integrated georeferencing and tachymetry amounts to 22 mm (see table 4.33), which is not significantly larger than the values achieved in study 1. To sum up, based on arbitrary images captured by our portable mobile mapping system that are processed using our integrated georeferencing approach, absolute 3D point coordinates can be computed with an accuracy of approx. 2 cm.

Camera head	# CPs	Δx [mm]	Δy [mm]	Δz [mm]	Δ3D [mm]
cam3 and cam4	13	10	14	12	22

Table 4.33: RMSE values of study 2 for check point residuals between integrated georeferencing and tachymetry.

Check Point Investigations for Building Dataset Muttenz18

We defined 36 points in total and used 28 of them as check points and 8 as ground control points (see figure 4.32). At least four image observations per point were provided for 3D coordinate computation, which was performed for both lidar SLAM-based camera poses and improved camera poses by integrated georeferencing. While the resulting mean check point precision by forward intersection amounts to 123 mm for lidar SLAM-based poses, it is 3 mm for improved poses by integrated georeferencing (see table 4.34). Please note that precision is an indicator for the relative accuracy of typical 3D distance measurement tasks.

	Lidar SLAM-based poses		Improved poses by integrated georeferencing	
	8 GCPs	28 CPs	8 GCPs	28 CPs
Precision [mm]	82	123	2	**3**
Accuracy [mm]	106	133	13	**18**

Table 4.34: Precision and accuracy values for ground control points and check points from both lidar SLAM-based camera poses and integrated georeferencing. Precision indicates the 3D RMSE of forward intersection for single point measurements. Accuracy shows the RMSE of 3D point residuals to tachymetry.

In order to determine accuracy values, we transformed the computed 3D points into the ground truth coordinate frame by eight GCPs (see figure 4.32). Hence, three translation and three rotation parameters were estimated, but no scale. Calculation of deviations between tachymetry and the transformed 3D points resulted in check point 3D RMSE values of 133 mm for lidar SLAM-based poses and 18 mm for improved poses by integrated georeferencing. The results of integrated georeferencing, which represent the absolute 3D measurement accuracy, are similar to the results obtained for the dataset Muttenz17.

In summary, we obtained precision and accuracy values for lidar SLAM-based poses at the decimeter level. By performing a subsequent integrated georeferencing procedure, we improved these values by an order of magnitude, so that relative measurements within the sub-centimeter range and absolute measurements at the centimeter level are feasible.

Discussion

Integrated georeferencing exploiting fixed ROPs and using several GCPs located in each corner of an indoor space provided absolute 3D accuracies of approx. 2 cm as well as precision values at the millimeter level. Compared to initial solutions based on lidar SLAM, these values indicate an improvement by an order of magnitude. Nonetheless, such a performance is only possible when extracting DSP-SIFT instead of SIFT features, which can compensate for the poor textures encountered in indoor environments.

Summary

We performed comprehensive investigations utilizing our integrated georeferencing approach. Incorporation of precalibrated ROPs among cameras enabled to orient all intended images with high accuracy, robustness and efficiency. Moreover, our developed procedure can cope with diverse environments and varying camera configurations. Employing GCPs and constraining ROPs resulted in absolute 3D natural point accuracies of approx. 2 cm for indoor environments. Depending on the use case, accuracy values for outdoor environments are slightly larger and amount to a few centimeters. Nonetheless, the obtained check point accuracies frequently correspond to GCP accuracies. In case of GCP utilization, prior EOPs determined by both direct georeferencing using GNSS/INS filtering and lidar SLAM are suitable. However, initial trajectory deviations should remain within the sub-meter range.

Not relying on any GCPs necessitates precisely calibrated ROPs among cameras in order to obtain accurate results. We achieved horizontal accuracies of ca. 5 cm for the dataset Zug17 and up to ca. 30 cm for the dataset Basel15. However, since the height component is often highly dependent on direct georeferencing solutions, these 2D values significantly increased for the 3D case. Nonetheless, incorporation of at least one GCP is particularly advantageous for the height component, allowing for absolute 3D accuracies within the sub-decimeter range. We further showed that self-calibration of ROPs among stereo camera systems by solely exploiting EOPs and relative poses of stereo bases as metric information is not only feasible but also precise.

Sophisticated computation of accurate camera poses constitutes the main part of this thesis. Once georeferenced, imagery captured by multi-camera systems can be exploited for the reconstruction of dense 3D scenes. Within the next chapter, we show that image orientations provided by our integrated georeferencing approach meet the requirements of sub-pixel accuracy, enabling precise and dense depth map generation. Furthermore, several configurations for in-sequence dense image matching are evaluated.

Chapter 5

Evaluation of In-Sequence Dense Image Matching

In order to compute dense 3D scene representations, camera pose estimation is followed by dense image matching (DIM) (see section 2.3). Since entirely dependent on the previous steps, not only precise calibration but also image orientations need to be available at the sub-pixel accuracy level. Hence, we evaluate DIM results to draw a conclusion about the quality of camera poses, but also to investigate various in-sequence matching scenarios for precise depth map generation. However, the main focus of this thesis is on integrated georeferencing, so that only the street-based dataset Basel14 (see section 4.2.1) is used for these experiments. Furthermore, most investigations solely employ stereo images captured by the forward pointing stereo camera system. Nonetheless, the densely built-up urban environment poses several challenges such as illumination changes and multiple occlusions, as well as large scale variations due to a higher depth of field. Some of these issues can be addressed by exploiting the high redundancy provided by multi-camera mobile mapping systems. In order to still allow efficient data processing, adequate image combinations have to be selected (see section 5.2). Neighboring images feature high similarity and low to medium scale differences of specific regions, which is beneficial for dense image matching. Potential configurations also consist of stereo pairs captured from cameras pointing in moving direction at different timestamps, which demands an alternative polar rectification method. Investigations in both image and object space show the potential of in-sequence dense image matching in terms of accuracy, reliability and completeness.

5.1 Precise and Dense Depth Map Generation for Image-Based Mobile Mapping

We aim at obtaining depth maps of high quality based on image sequences captured by mobile mapping systems. Once distortion-corrected images as well as corresponding camera poses by integrated georeferencing are available, image rectification needs to be performed (see figure 5.1). Resulting epipolar images facilitate the subsequent step of dense image matching by reducing the image correspondence problem to a 1D issue. Based on depth maps, dense 3D point clouds are computed, which can be fused and filtered. Reprojection of this refined 3D geoinformation by incorporation of the viewing geometry again leads to depth maps that enable comparisons.

Accurate, reliable and complete depth maps are best generated by exploiting the available image redundancy. Processing stereo images recorded at the same point of time is a standard procedure that works well. However, rectifying image pairs acquired at different epochs by cameras predominantly oriented in moving direction poses new challenges. Epipoles located inside or close to the images do not allow for appropriate in-sequence dense image matching (see figure 5.2 and figure 5.3). Hence, an advanced rectification method is required. We use the multi-view stereo matching software SURE (Rothermel et al., 2012) for our experiments. In order to overcome strong motion in viewing direction,

it features a polar rectification method. Its implementation is presented within the following section, as well as in more detail in Cavegn et al. (2015) and Rothermel (2017).

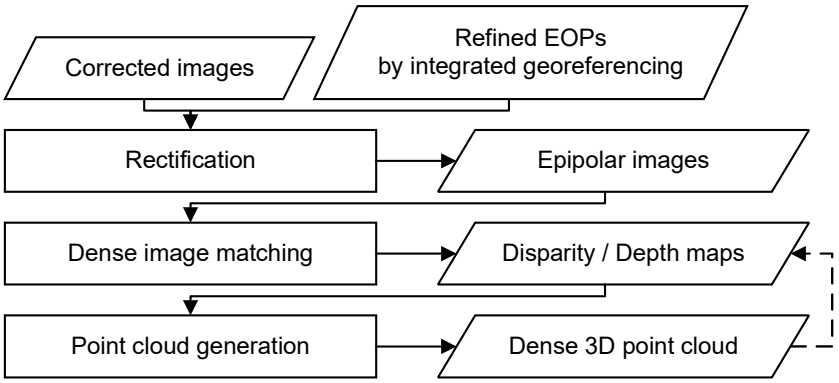

Figure 5.1: Computation of 3D geometry based on oriented images.

Polar Rectification Approach

Standard rectification approaches do not work for arbitrary geometric configurations of stereo image pairs. Fusiello et al. (2000) as well as Loop and Zhang (1999) construct virtual image planes parallel to the baseline connecting the two camera centers. Then homographic mapping is used to project the original images onto the virtual images planes. This is especially problematic for motion in viewing direction, since virtual and original image planes are close to perpendicular, which results in huge image dimensions as well as large distortions of the rectified imagery (see left part of figure 5.3). This does not only increase processing times, but makes dense matching challenging or even impossible as heavily distorted images cause problems in the computation of similarity measures.

A rectification method for general motion, also handling pure forward motion as present in our mobile mapping scenarios, was proposed by Pollefeys et al. (1999). The SURE implementation follows their approach, where rectification is performed by sampling corresponding half epipolar lines across two views. Hence, half epipolar lines are the line segments defined by subdividing the epipolar line at the epipole (see figure 5.2). Half epipolar lines are subsequently processed in a circular scheme. Corresponding lines in the images are then arranged in parallel image rows in the rectified images $\mathbf{I}_{r,\theta}$.

The dimensions of the resulting images are limited by enforcing the distances of subsequent epipolar lines such that each pixel on the image border opposing the epipole is sampled exactly once. Moreover, this adaptive sampling implicitly avoids the occurrence of pixel compression in $\mathbf{I}_{r,\theta}$, but necessitates a lookup table to invert the transformation. Epipoles nearby or inside the images cause strongly distorted image regions, where depth determination is inherently not accurate (Pollefeys et al., 1999). In order to avoid the influence of this singularity, the affected image region A_d is automatically detected and discarded from further processing.

Following the rectification step, dense image matching is carried out on $\mathbf{I}_{r,\theta}$. Interpolation in a lookup table, which is established during the rectification process, enables determining the corresponding Cartesian coordinates u, v from polar coordinates r, θ. Subsequently, SURE performs structure computation, which is described in detail by Cavegn et al. (2015).

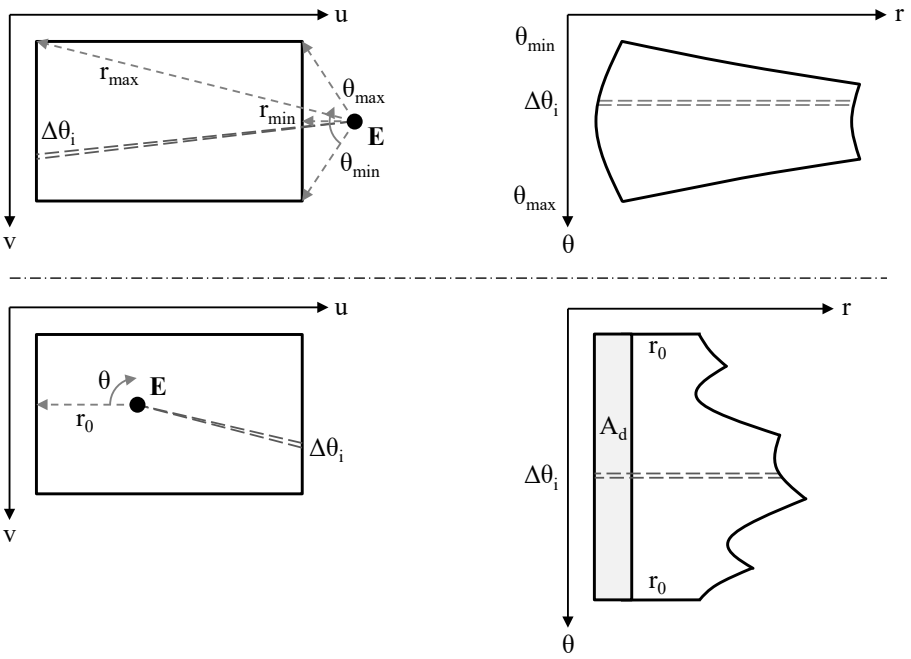

Figure 5.2: Polar rectification process. Transformation of $\mathbf{I}_{u,v}$ (left) to $\mathbf{I}_{r,\theta}$ (right) for the case the epipole E is located outside the image (top) and for the case the epipole E is located inside the image (bottom). A_d is the distorted image region that will be removed for depth estimation.

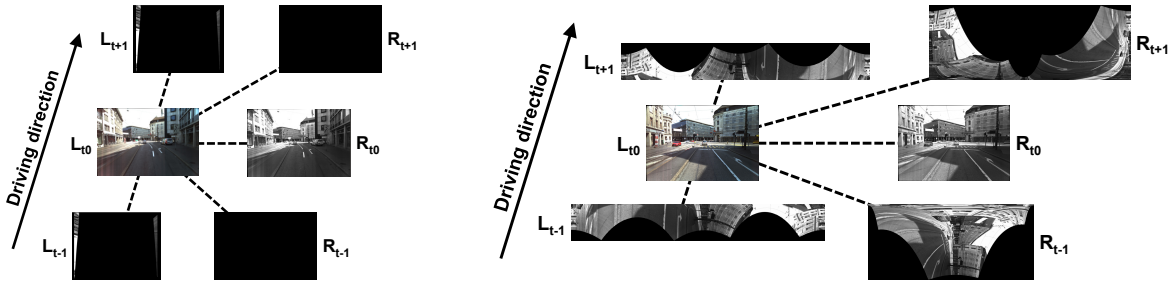

Figure 5.3: Base image (L_{t0}) and its five neighboring images rectified by the SURE standard procedure according to Fusiello et al. (2000) (left) as well as rectified by SURE using polar rectification (right).

5.2 Configurations for Dense Image Matching

Stereo camera systems pointing in the driving direction as depicted in figure 4.1 typically provide highly redundant imagery. In order to exploit this redundancy, images captured at different epochs need to be incorporated into the dense image matching process. Hence, we selected the four matching configurations c1 to c4 depicted in figure 5.4. Configuration c1 represents standard stereo matching with one base and one match image captured at the same point of time. Numerous sophisticated algorithms exist for two-view matching (Scharstein and Szeliski, 2002; Menze and Geiger, 2015). Configuration c2 stands for the case in which only mono imagery would be available. It is limited to sequential matching of the base image with the previous and the following image. This case puts high demands on providing sufficient relative orientation accuracies, but is not dependent on precise synchronization of multiple cameras. With configurations c3 and c4 we introduce and investigate two multi-view stereo approaches, for which qualitatively better results can be expected. In case of configuration c4, the base image is matched with

all five neighboring images. Omitting the two match images of configuration c2 from configuration c4 leads to configuration c3.

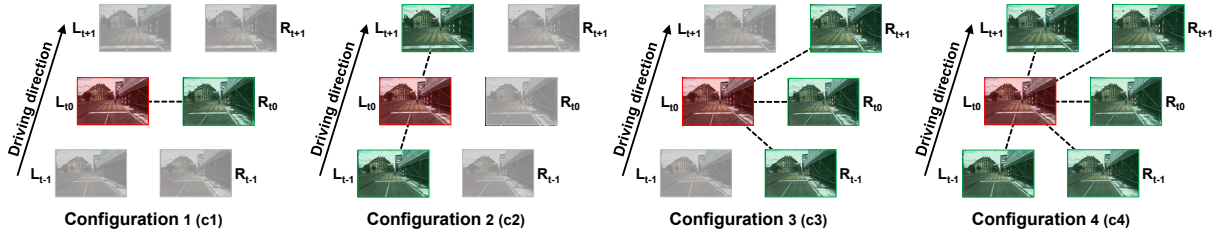

Figure 5.4: Selected image matching configurations, red: base image, green: match images.

5.3 Point Cloud Generation and Filtering

We aimed to generate accurate, reliable and dense 3D point clouds based on the three image sequences of the dataset Basel14 that are depicted in figure 5.5. Hence, we processed this mobile mapping imagery using the software system SURE. It turned out that processing with standard parameters already leads to good results. Potential fine-tuning does not cause a significant improvement, while filtering in object space using the octree-based approach described in Wenzel et al. (2014) is crucial. First, we compared the impact of using fold 3 instead of fold 2 in the filtering procedure (see figure 5.6). This means that each point in object space needs to be confirmed by two and not only by one additional point. Obviously, it resulted in a lower density, thus causing less clutter especially around overhead wires and eliminating most points representing moving objects, e.g. a streetcar in the middle of the left part of figure 5.6.

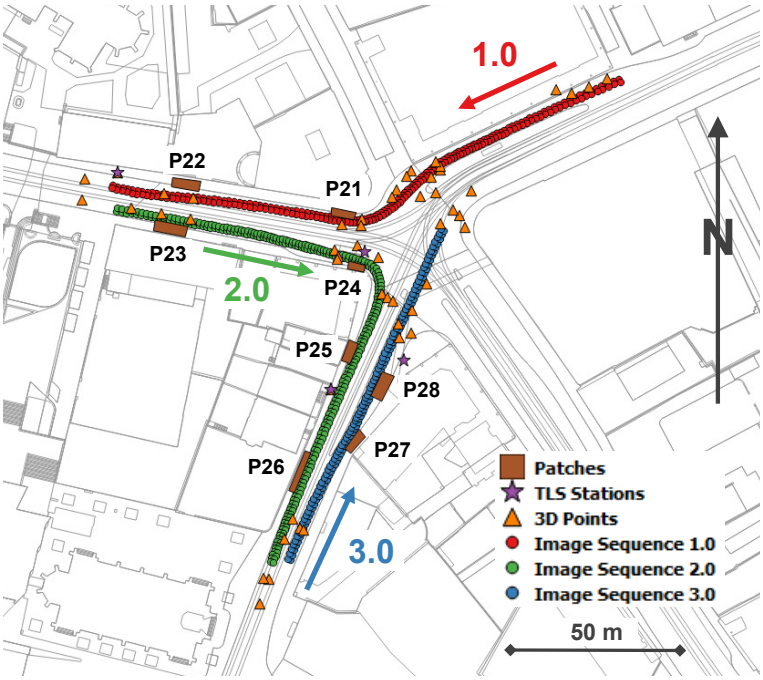

Figure 5.5: Base map of the study area with overlaid projection centers of selected stereo image sequences, 3D reference points, terrestrial laser scanning (TLS) stations and point cloud patches (Source of background map: Geodaten Kanton Basel-Stadt).

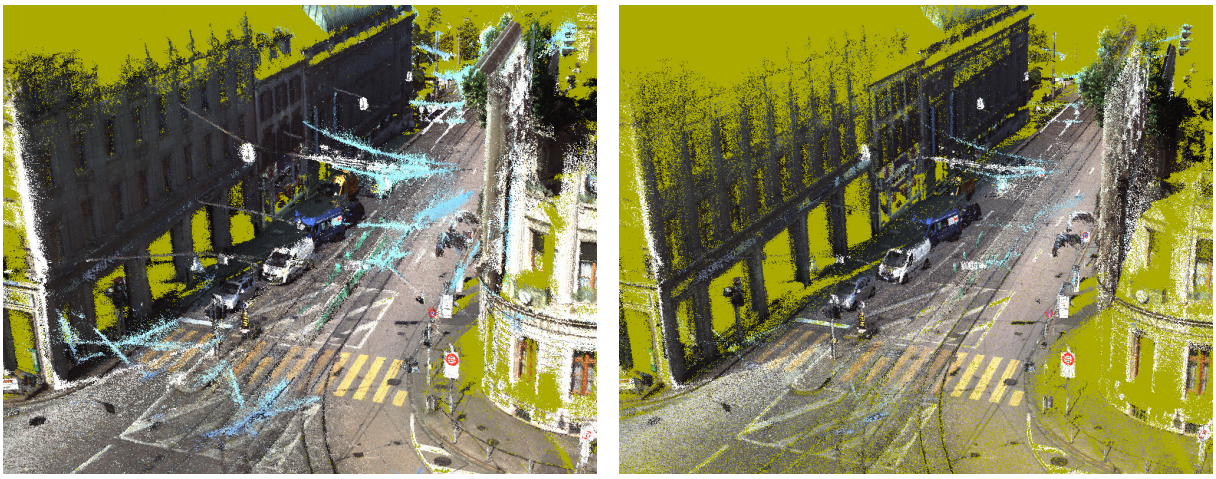

Figure 5.6: Filtered point clouds generated by incorporating forward stereo imagery of sequences 1.0 and 2.0. Exploitation of five match images per base image (c4) using SURE filtering fold 2 (left) and SURE filtering fold 3 (right).

Second, we filtered SURE point clouds in object space using fold 3 and investigated the results of configurations c1 and c4 (see figure 5.7). While almost all moving objects are eliminated when incorporating five match images (see right part of figure 5.7), several parts of moving objects remain if just one stereo pair is matched (see left part of figure 5.7). Only performing standard stereo matching leads to more clutter and a considerable number of sky points around overhead wires. The fact that more façades are mapped in the right part of figure 5.7 is mainly due to the minimum forward intersection angle of 2° for both configurations.

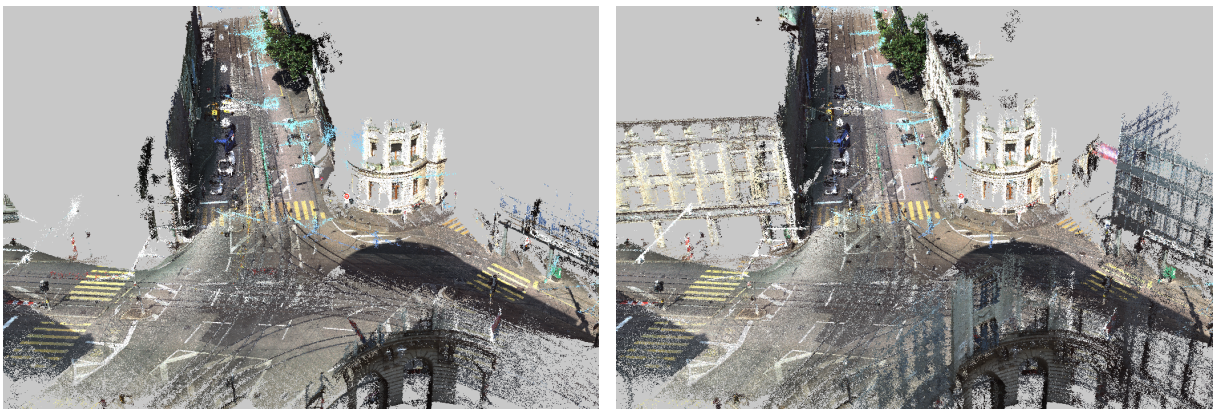

Figure 5.7: Filtered point clouds generated by incorporating forward stereo imagery of sequences 1.0, 2.0 and 3.0. Exploitation of one match image per base image (c1, left) and five match images per base image (c4, right) using SURE filtering fold 3.

In order to enable the forward stereovision system to particularly generate more points on sidewalks and façades, we set the filtering parameter fold to 2. Furthermore, this allowed to demonstrate the benefit of additionally exploiting imagery from the back-right as well as from the left stereovision system. While employing back-right imagery leads to a significant increase in sidewalk points (middle bottom part of figure 5.8), the left stereovision system covers a larger road surface part and it is beneficial for lower façade points as well (right bottom part of figure 5.8).

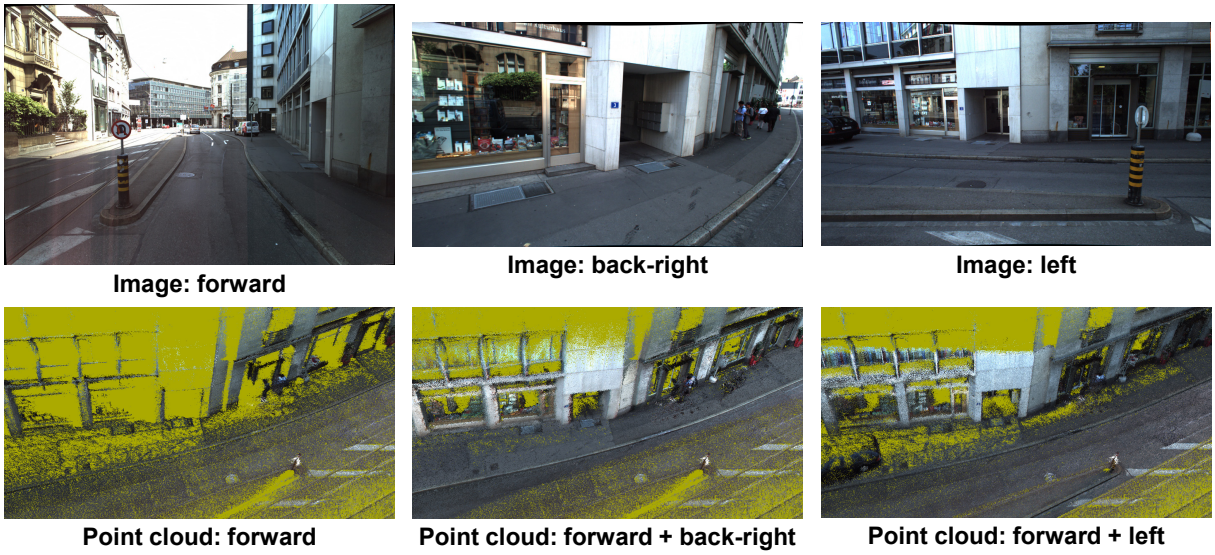

Figure 5.8: Mobile mapping images and generated point clouds by configuration c4 in the region of patch P27 (see figure 5.5) using SURE filtering fold 2.

5.4 Investigations in Image Space

Accurate depth maps are fundamental for the urban model of 3D image spaces (Nebiker et al., 2015) and in particular for reliable and accurate 3D monoplotting applications. Therefore, we performed comparisons of depth maps, which were either generated directly by the SURE triangulation module or obtained by back-projecting point clouds into the viewing geometry of a base image. Similar to the methodology of Geiger et al. (2012), who did not interpolate ground truth disparity maps in order to avoid artificial errors, we did not interpolate depth maps, either. This allowed to evaluate the raw depth values and to cope with missing parts of depth maps. We only computed depth deviations for pixels holding values for both depth maps and we just considered deviations smaller than 50 cm for RMSE and mean calculation.

Differences to Configuration c4

In a first series of tests, we carried out relative depth comparisons in image space with the depth map of configuration c4 as reference (see figure 5.9 and table 5.1). For all extracted 3D base images, c1-c4 delivered the lowest RMSE values. The largest RMSE as well as mean values were computed for c2-c4. While RMSE values for c1-c4 and c3-c4 are in the range of 36 mm to 57 mm, the range for c2-c4 is from 57 mm to 79 mm. c3-c4 delivered the most façade points and c2-c4 shows an opposite behavior compared to the two other configurations with inverted depth differences. In c2-c4 and c3-c4 the region close to the epipole, where depth estimations are not accurate and thus removed, is clearly visible. This effect results in a considerable number of road surface points not being mapped.

	Base image 1.0.1		Base image 2.0.1		Base image 2.0.2		Base image 3.0.1	
	RMSE [mm]	Mean [mm]	RMSE [mm]	Mean [mm]	RMSE [mm]	Mean [mm]	RMSE [mm]	Mean [mm]
c1-c4	53	-7	36	-10	53	-6	55	-11
c2-c4	72	9	57	22	64	19	79	18
c3-c4	56	4	38	1	54	3	57	-3

Table 5.1: Numerical deviations of depth maps generated by the SURE triangulation module. Depth maps of configuration c4 serve as reference. Base image numbering is as follows: the first digit corresponds to the street section (see figure 5.5), the second indicates the campaign, and the third shows a consecutive numbering.

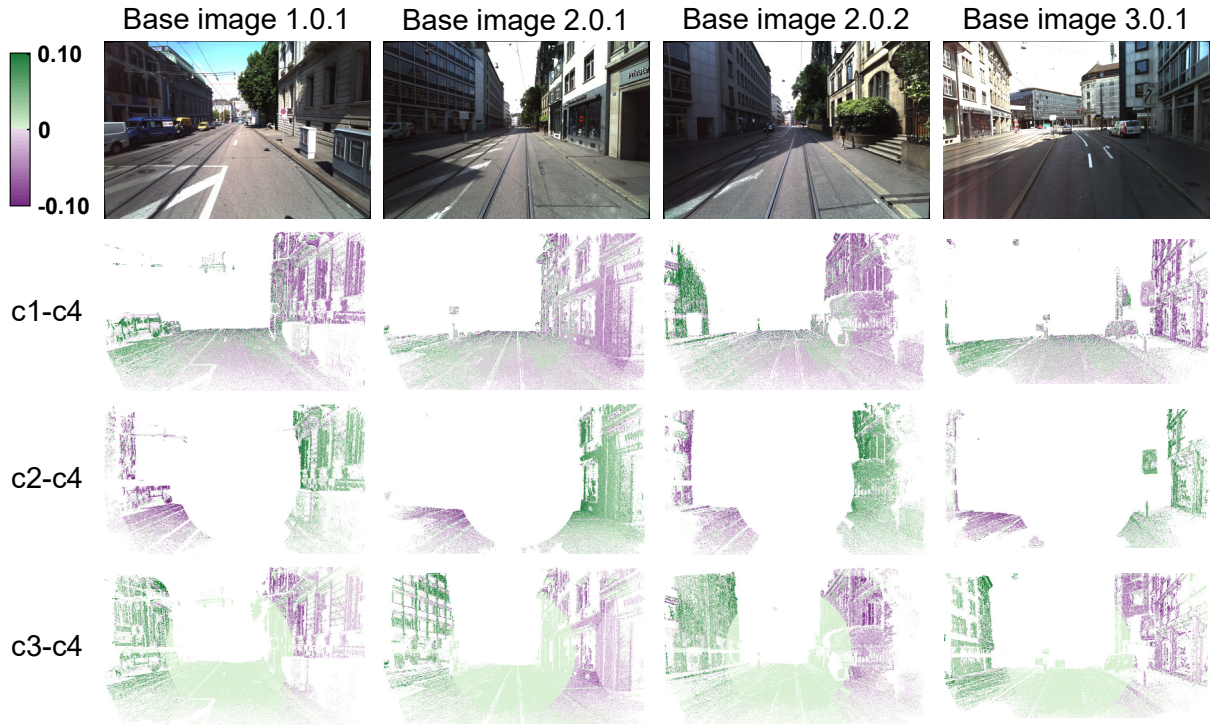

Figure 5.9: Visual deviations of depth maps generated by the SURE triangulation module. Depth maps of configuration c4 serve as reference. Base image numbering is as follows: the first digit corresponds to the street section (see figure 5.5), the second indicates the campaign, and the third shows a consecutive numbering.

Differences to a TLS Reference

In a second test series, we used terrestrial laser scanning points projected into image space as reference. We compared these reference depth maps (TLS) with dense image matching point clouds for all matching configurations (c1-c4), all generated by the SURE triangulation module. Whereas we used fold 1 for configuration c1, we set the fold parameter to 2 for the other configurations. We computed the largest mean values, i.e. the largest depth offsets, for base image 1.0.1 (see table 5.2). We observed the largest RMSE, i.e. the largest depth noise values, for base images 2.0.2 and 3.0.1, which were caused by distinctive shadow areas and vegetation (see figure 5.10). All RMSE values are in the range of 107 mm to 174 mm, with the lower bound featured by c1-TLS. RMSE values for c4-TLS are slightly larger than for c3-TLS, but c4-TLS also includes more depth observations.

	Base image 1.0.1		Base image 2.0.1		Base image 2.0.2		Base image 3.0.1	
	RMSE [mm]	Mean [mm]	RMSE [mm]	Mean [mm]	RMSE [mm]	Mean [mm]	RMSE [mm]	Mean [mm]
c1-TLS	107	30	121	7	145	-3	145	-17
c2-TLS	153	18	127	5	160	-6	174	-8
c3-TLS	137	15	142	-4	164	-16	146	-10
c4-TLS	144	8	143	-4	169	-11	150	-8

Table 5.2: Numerical depth deviations between point clouds generated by the SURE triangulation module and TLS. Base image numbering is as follows: the first digit corresponds to the street section (see figure 5.5), the second indicates the campaign, and the third shows a consecutive numbering.

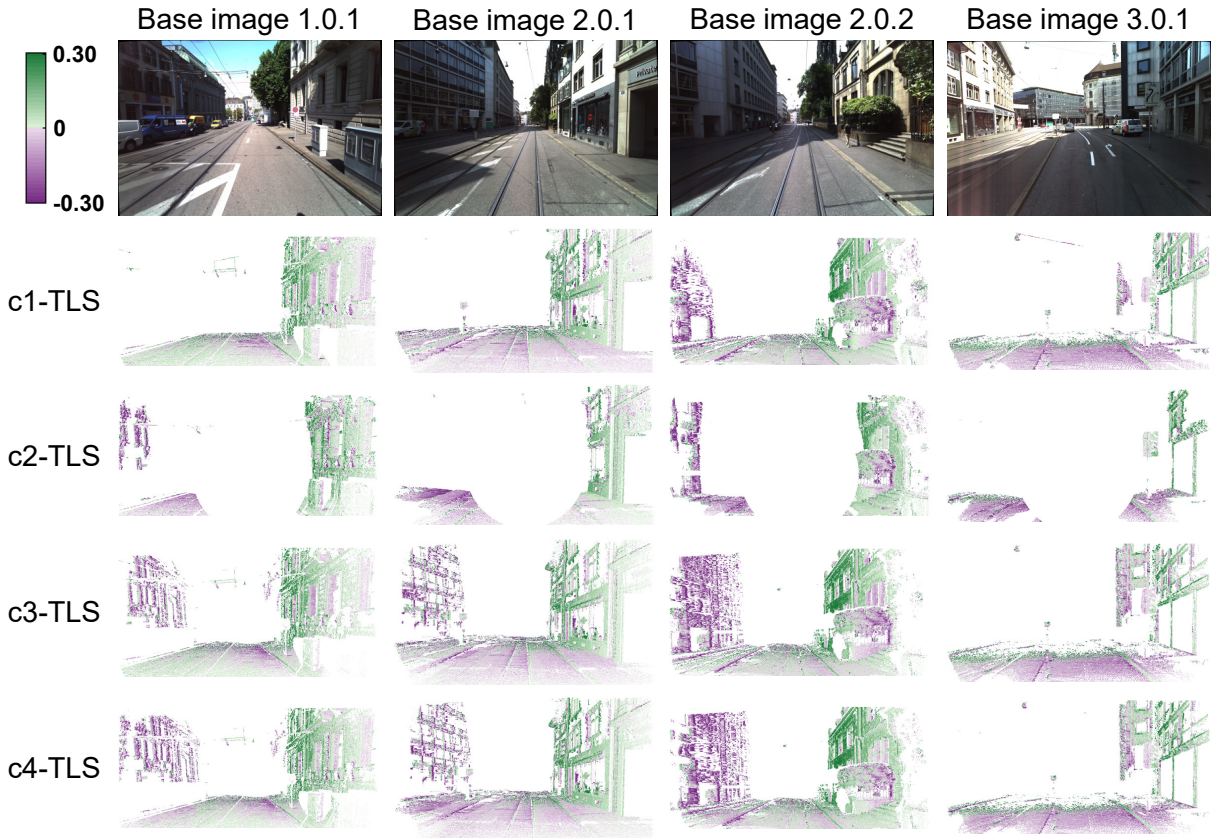

Figure 5.10: Visual depth deviations between point clouds generated by the SURE triangulation module and TLS. Base image numbering is as follows: the first digit corresponds to the street section (see figure 5.5), the second indicates the campaign, and the third shows a consecutive numbering.

Discussion

Investigations in image space showed similar results between the four selected configurations for all four base images. Due to a single large stereo base, the traditional stereo configuration c1 provided high accuracies, which did not further improve with the additional use of images captured at different epochs. Configuration c2 with sequential matching of mono image sequences loses many points around the epipole. It also yields a limited accuracy since the base for image ray intersection is very small. The differences between configurations c3 and c4 are not significant. However, especially compared to the standard stereo configuration c1, the increasing number of match images available in configuration c4 delivers significantly higher point densities. In summary, the traditional stereo matching configuration c1 delivers depth maps with a medium completeness but with the highest accuracy, and configuration c4 yields depth maps with the highest completeness but slightly larger RMSE values.

5.5 Investigations in Object Space

Visual inspections give a first impression of differences between selected point clouds. However, in order to be able to draw meaningful conclusions in terms of completeness and accuracy, assumptions need to be verified by numerical values. According to section 2.3.2, density and deviation values for several patches were computed and visualized using deviation patches and profiles. We decided to analyze point

cloud patches, even though point clouds of complex 3D structures can be generated by image-based mobile mapping. However, their evaluation with terrestrial laser scanning is aggravated due to different viewpoints and measuring techniques.

Selection and Processing of Patches

First of all, we selected five road (P1-P5) and six façade patches (P11-P16) that are predominantly planar and we defined them arbitrarily by four corners. Samples for each of the eleven patches depicted in Cavegn et al. (2015) show that all road patches contain road markings and have varying lighting conditions due to shadows, all but P5 feature streetcar rails, P3 and P4 include curbstones. Façade patches contain façade parts of different structure and material. Later on, we additionally determined eight patches (P21-P28) in road and sidewalk regions by four points, i.e. two patches per side of the roads that were mapped in both directions (see figure 5.5). Each patch area needed to be covered with point clouds generated by all the three stereovision systems forward, back-right and left as well as with points captured by TLS. No disturbing objects were present in most cases and all patches include curbstones as well as one vertical plane of them. Since all patch borders entirely lying on the road surface are defined by the back-right stereovision system, the limitation on the opposite side is given by vertical objects like façades or walls. As it can be seen in figure 5.11, we chose patches with varying illumination conditions and a different portion of road markings.

In order to assess all 19 selected patches, we used the evaluation procedure described by Cavegn et al. (2014). First, we extracted TLS and DIM point clouds for each patch (see top left part of figure 5.12). For patches P1-P5 and P11-P16, extraction was performed from one of the three filtered and fused point clouds generated by forward stereo images, i.e. either from image sequence 1.0, 2.0 or 3.0. For patches P21-P28, we additionally incorporated point clouds from the back-right and left stereovision systems. Second, all reference TLS point cloud patches were subsampled to a distance of 1 cm. Furthermore, we used TLS and DIM grids of 3 cm spacing for P21-P28 and grids of 5 cm spacing for the other patches in order to compute deviations.

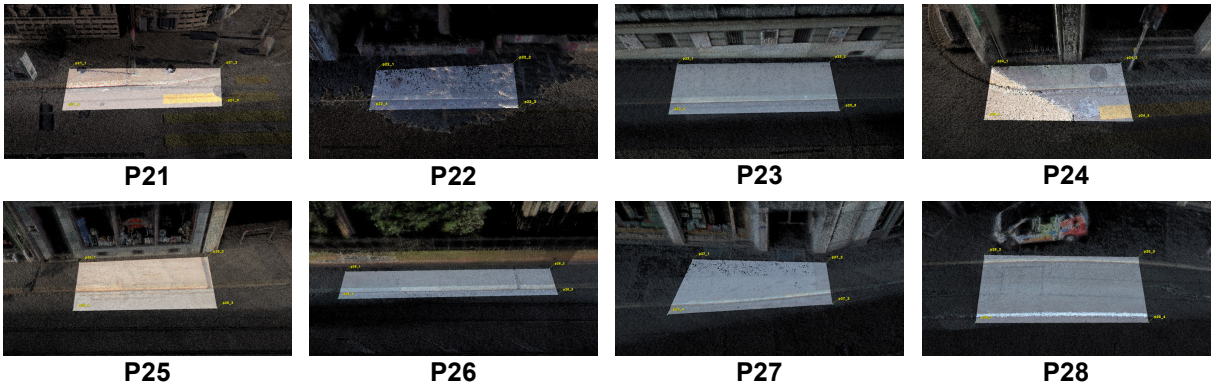

Figure 5.11: Selected segments of point clouds generated by forward and back-right stereo imagery defining the patches P21-P28 (see figure 5.5).

Deviation Patches and Profiles

The bottom left part of figure 5.12 exemplarily shows the deviations of road patch P2 for configurations c1 and c4. Differences are not significant, and rails are clearly indicated by positive deviations while road surface deviations are mainly negative. Cross-track and along-track profiles reveal an offset of about 1 cm for both configurations compared to TLS and there is a little less noise for configuration c4 (see right part of figure 5.12).

The left column of figure 5.13 illustrates the deviations of patch P27 for configurations c1 and c4 as well as for all three stereovision systems. White holes indicate sparse regions where deviations were not computed. Deviations for forward and back-right are mainly positive and the largest deviations are

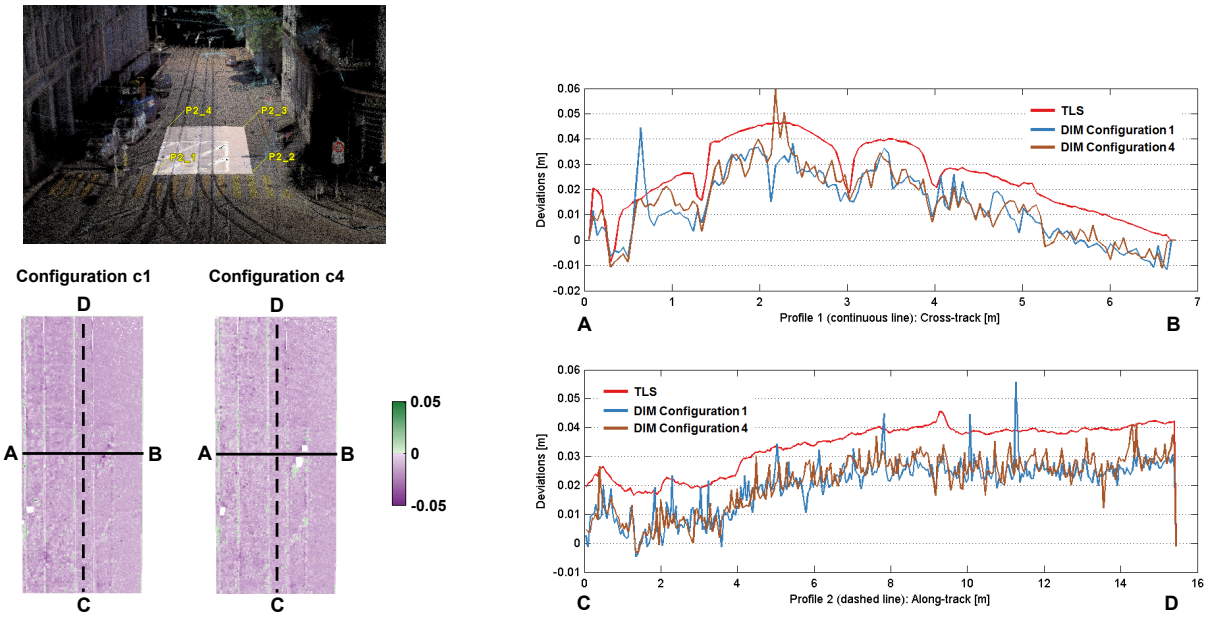

Figure 5.12: Extraction of road patch P2 (top left), corresponding deviations DIM-TLS (bottom left) as well as cross-track and along-track profiles (right).

computed along curbstone edges of the sidewalk as well as on the upper part of the sidewalk. There are significantly more points for back-right than for forward, especially in case of c4. Profiles of patch P27 are depicted by the right column of figure 5.13. Cross-track and along-track profiles of point clouds from image sequence 2.0 (forward and back-right) reveal an offset of ca. 2 cm for both configurations compared to TLS. The vertical curbstone plane is the less accurately modeled by the back-right stereovision system. Point clouds of c4 are less noisy than point clouds of c1 for forward and left.

Road and Façade Patches: c1 vs. c4

Numerical values for the road patches P1-P5 are given in table 5.3 and statistics of the façade patches P11-P16 are listed in table 5.4. While the size of road patches ranges from 82 to 138 m^2, the size of façade patches lies between 23 and 119 m^2. Configuration c1 delivers significantly more points than configuration c4 for road patches, which is due to the fold parameter for the SURE triangulation module (fold 1 for c1 and fold 2 for c4, but fold 3 for both configurations in the subsequent filtering step). In contrast, all façade patches but patch P12 have a higher density for configuration c4 than for configuration c1. The reason for this is a minimum angle value of 2° for the SURE triangulation module for both configurations. Low density values are caused by difficult matching conditions, which is a large shadow area in patch P3 and a high number of road marking points whose pixel information is partly overexposed in patch P5. RMSE, mean and standard deviation values are almost identical for configurations c1 and c4 in case of road patches. Patch P16 features significantly more points since this is the only investigated façade which is almost perpendicular to the driving direction. Mean and thus RMSE values are larger for configuration c4 than for configuration c1, but there is the same standard deviation value. The larger values of patch P15 are due to different façade levels and non-planarity. Mean standard deviation values for road patches are approx. 1 cm and around 5 times larger for façade patches.

Forward, Back-Right and Left Stereovision Systems: c1 vs. c4

Mean density and deviation values for the mixed road and sidewalk patches P21-P28 are given in table 5.5, and standard deviation values are depicted in more detail by figure 5.14. While the size of the investigated patches ranges from 11 to 32 m^2, the mean size of 22 m^2 corresponds to around one fifth of the mean value of the road patches P1-P5. The highest mean density value of 23268 points/m^2 was computed for

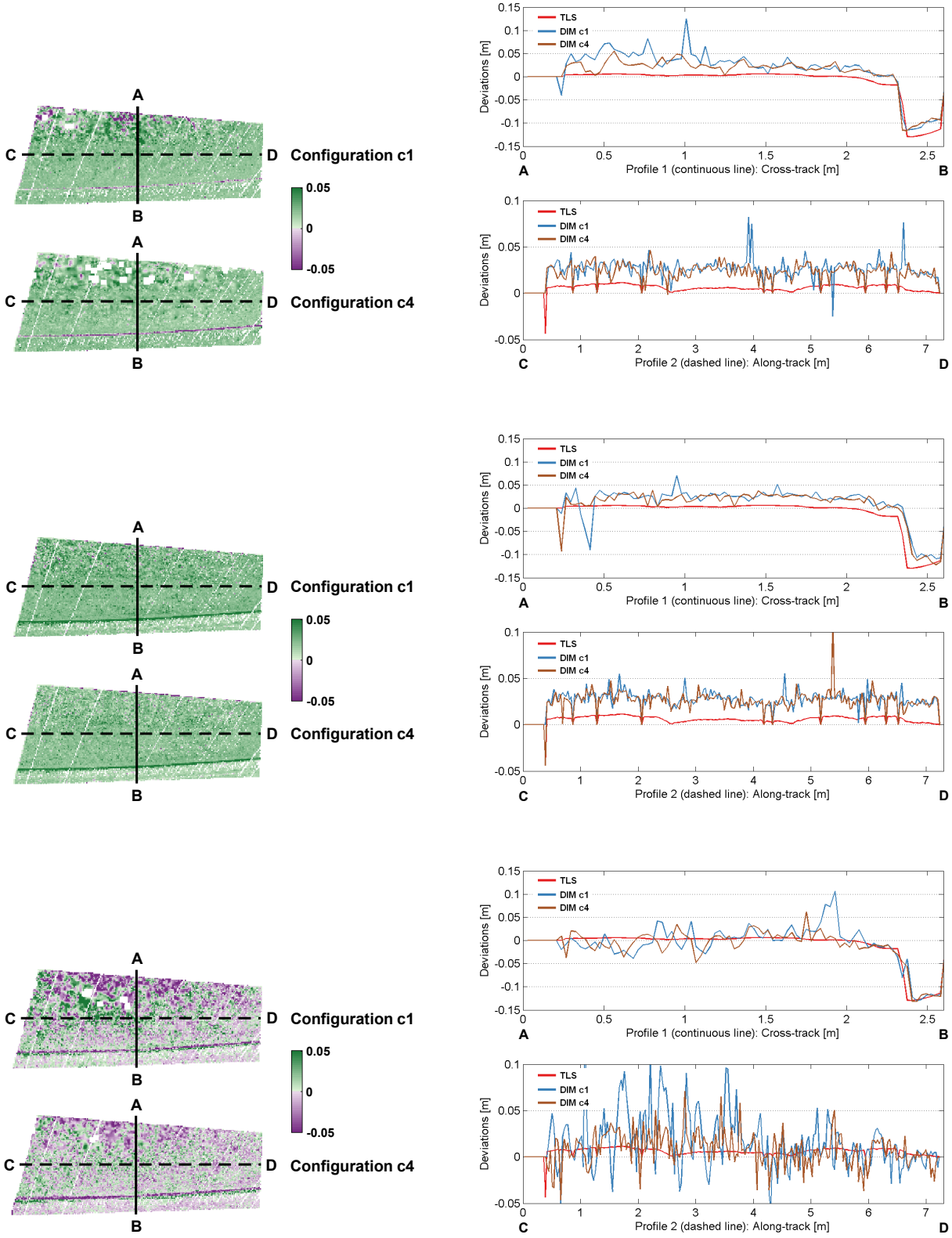

Figure 5.13: Deviations DIM-TLS and profiles of road patch P27 for the forward (top), back-right (middle), and left (bottom) stereovision systems.

	Patch size [m^2]	Density [Points/m^2]	RMSE DIM-TLS [mm]	Mean DIM-TLS [mm]	SD DIM-TLS [mm]
P1 fw c1	103	1742	7	-6	5
P1 fw c4	103	900	7	-5	5
P2 fw c1	105	1466	13	-12	6
P2 fw c4	105	909	12	-11	6
P3 fw c1	82	692	22	-20	9
P3 fw c4	82	318	20	-18	9
P4 fw c1	90	1729	14	-5	13
P4 fw c4	90	1006	14	-3	14
P5 fw c1	138	1063	12	10	7
P5 fw c4	138	624	13	10	7
P1-P5 fw c1	104	1338	14	-7	8
P1-P5 fw c4	104	751	13	-5	8

Table 5.3: Density and deviation values of all road patches for configurations c1 and c4 using SURE filtering fold 3.

	Patch size [m^2]	Density [Points/m^2]	RMSE DIM-TLS [mm]	Mean DIM-TLS [mm]	SD DIM-TLS [mm]
P11 fw c1	81	1166	47	32	34
P11 fw c4	81	1340	65	51	40
P12 fw c1	61	729	57	46	33
P12 fw c4	61	701	74	66	33
P13 fw c1	119	908	72	-6	72
P13 fw c4	119	1152	75	20	72
P14 fw c1	38	1006	74	-68	30
P14 fw c4	38	1132	88	-81	34
P15 fw c1	85	606	89	-63	63
P15 fw c4	85	1113	121	-109	54
P16 fw c1	23	1580	67	-12	65
P16 fw c4	23	2106	79	45	65
P11-P16 fw c1	68	999	68	-12	50
P11-P16 fw c4	68	1257	84	-1	50

Table 5.4: Density and deviation values of all façade patches for configurations c1 and c4 using SURE filtering fold 3.

forward c1, which is circa three times higher than for forward c4, principally caused by the fold parameter for the SURE triangulation module (fold 1 for c1 and fold 2 for c4, but fold 2 for both configurations in the subsequent filtering step). Both values are 10-20 times higher than the values determined for P1-P5, which is mainly due to a different filtering degree (fold 3 for P1-P5 and fold 2 for P21-P28). Low density values are caused by difficult matching conditions such as a large shadow area for patch P22 in case of the forward and back-right stereovision systems resulting in rather high standard deviation values, but not in case of the left stereovision system where just a small shadow area is present since captured at another date and daytime.

In most cases, RMSE and standard deviation values are larger for c1 than for c4. Because of the lower filtering degree and a vertical plane of curbstones for each patch, standard deviation values are approx. twice as large as for road patches P1-P5. While mean DIM-TLS values of 13-23 mm were computed for patches P27 and P28 for forward and back-right, there are just a few millimeters for the left stereovision system whose data was captured from another trajectory (image sequence 2.0 vs. 3.0). The largest standard deviation value for forward c1 was determined for patch P24, which is due to a bicyclist who caused many non-road points that could not be removed by c1 but almost completely by c4.

3D point accuracy depends on the measuring distance that is depicted by Burkhard et al. (2012) for both camera types used in the present investigations. As the average distance between the left stereovision system and the selected patches is around 10 m, the larger accuracy values compared to the forward and back-right stereovision systems is not surprising. While patch P26 shows the best values for the left stereovision system with a distance of 7 m, patch P24 has the second largest standard deviation values due to a distance of 13 m. The largest RMSE and standard deviation values were computed for patch P23 which has the largest area.

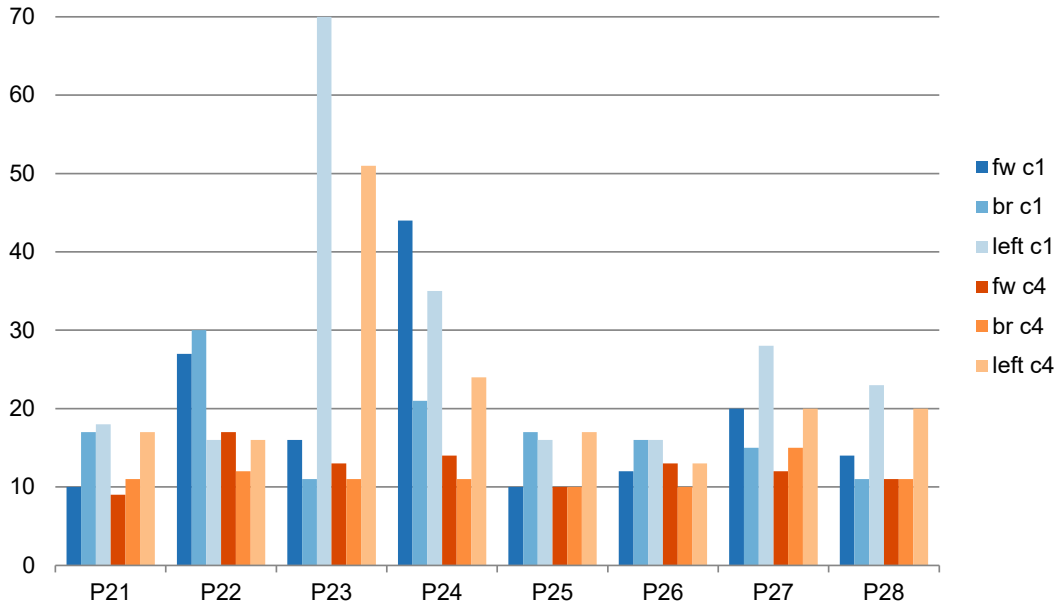

Figure 5.14: Standard deviation values in mm for residuals between DIM and TLS point cloud patches (see SD DIM-TLS in table 5.5) (fw: forward, br: back-right, c1: stereo matching, c4: in-sequence stereo matching).

	Patch size [m^2]	Density [Points/m^2]	RMSE DIM-TLS [mm]	Mean DIM-TLS [mm]	SD DIM-TLS [mm]
P21-P28 fw c1	22	23268	26	-7	19
P21-P28 fw c4	22	7623	22	-9	12
P21-P28 br c1	22	16639	26	-7	17
P21-P28 br c4	22	15699	22	-9	11
P21-P28 left c1	22	3884	33	-13	28
P21-P28 left c4	22	4183	27	-13	22
P1-P5 fw c1	104	1338	14	-7	8
P1-P5 fw c4	104	751	13	-5	8

Table 5.5: Mean density and deviation values of all road patches using SURE filtering fold 2 (fw: forward, br: back-right, c1: stereo matching, c4: in-sequence stereo matching).

Discussion

In terms of density, large values were computed for both configurations c1 and c4 for the back-right stereovision system. In contrast, the forward stereovision system shows density values that are approximately three times larger for c1 than for c4. However, density is highly dependent on filtering parameters and especially on the filtering degree. In contrast to façade patches, we computed slightly more accurate values for c4 than for c1 in the case of road patches, which is caused by the higher redundancy of c4. Mainly due to large distances between the stereo cameras and the respective patches, the left stereovision system provides the less accurate point cloud patches. Similar accuracies were determined for the back-right and forward patches, since the lower resolution of the back-right compared to the forward stereovision system is compensated by shorter distances to the patches.

Summary

Mobile mapping scenarios often provide highly redundant imagery, making the selection of suitable image combinations crucial during multi-view stereo. Hence, we investigated the impact of different stereo and image sequence matching strategies on the geometric quality of both extracted depth maps and generated point clouds. To this end, we utilized the SURE software system with its implemented polar rectification approach. Subsequent steps of the SURE pipeline include dense image matching, triangulation, point cloud filtering and fusion. Especially for the triangulation and filtering modules, varying parameters were used (fold 1 or 2 and fold 2 or 3, respectively). Standard stereo matching led to high accuracies, which could not significantly be improved by additional in-sequence matching. However, the redundancy achieved by incorporating imagery from additional epochs into the dense image matching process resulted in more complete and reliable depth maps and point clouds.

Chapter 6

Conclusion and Outlook

6.1 Summary

Our Integrated Georeferencing Approach Exploiting EOPs, GCPs and ROPs

Our developed georeferencing approach provides accurate and reliable image orientations, while being efficient and versatile. The main reasons for these achievements are exploiting high image redundancies and constraining relative orientation parameters (ROPs) among cameras. To this end, multi-view image sequences captured by camera-based mobile mapping systems need to be available. In order to accurately georeference imagery in a predefined coordinate reference frame, we rely on initial exterior orientation parameters (EOPs) and optionally on ground control points (GCPs). These prior EOPs also allow for direct triangulation of all scene points within our global SfM pipeline, which is much more efficient than an incremental procedure.

We evaluated our integrated georeferencing approach through extensive investigations using six different real-world datasets. By exploiting EOPs, GCPs and ROPs, the accuracy potential lies at the centimeter level for absolute 3D coordinates and at the millimeter level for relative 3D measurements. If no GCPs are employed, it is challenging to meet accuracy requirements of better than one decimeter in urban canyons. We obtained horizontal accuracies of a few centimeters for a scenario featuring some loops, while dropping down to a few decimeters for an extended junction area. Since the height component is even more dependent on prior EOPs from direct georeferencing, 3D accuracies derived from integrated georeferencing between one and several decimeters are the normal case in urban environments. However, using just one GCP enables the elimination of systematic effects, which results in 3D accuracies within the sub-decimeter range. Nevertheless, this minimal GCP configuration needs to be complemented by at least one check point in order to guarantee a reliable solution. Furthermore, precisely calibrated multi-camera rigs and thus a correct inner geometry of trajectories are presumed. In case of solely precise stereo bases, ROPs among individual stereo camera systems can be self-calibrated, leading to relative 3D point accuracies in object space at the centimeter level.

A cost-effective solution would require no GCPs. While this is possible with our integrated georeferencing approach, we recommend to use at least one GCP in urban environments, so that potential systematic errors can be removed. Precise calibration of multi-camera systems is demanding and time-consuming. Our developed integrated georeferencing procedure allows for self-calibration of both interior and relative orientation parameters. This is particularly advantageous if a multi-camera system needs to be assembled on-site, which often happens in railway missions. Minimal requirements are prior camera poses from direct georeferencing or SLAM, a precisely calibrated stereo base or a length reference bar for metric information, as well as overlapping mappings, e.g. image acquisition in opposite directions. We demonstrated that the estimation of relative orientations among individual stereo camera systems by fixing stereo bases works well for a train-based MMS. This is even more challenging than road environments that feature better scene structures and street-based MMS that are more dynamic in terms of vehicle motion. Moreover, our procedure is able to overcome difficult feature matching conditions as

encountered in rail and indoor environments. While versatile, our integrated georeferencing approach showed robustness by successfully orienting all intended images.

Comparison of our Integrated Georeferencing Approach with State of the Art

Similar to us, Fanta-Jende et al. (2019) aim at improving trajectories of street-based mobile mapping vehicles in challenging urban environments. While we often rely on ground control points, they exploit airborne images as reference data. First, they developed a comprehensive automated procedure for the co-registration of aerial nadir and panoramic mobile mapping imagery based on salient road markings. Even though Jende et al. (2018) showed significant improvements and the feasibility to obtain 3D accuracies at the decimeter level, this was not confirmed in Fanta-Jende et al. (2019). Instead, they encountered a slight accuracy decrease or just a minor improvement by employing nadir images. Second, Fanta-Jende et al. (2019) presented a co-registration approach based on aerial oblique images that uses building façades or other high vertical objects as correspondence regions. While they obtained a significant improvement for one test area, two other test areas showed hardly any impact (Fanta-Jende et al., 2019). Their combined and thus best performing solution utilizing both nadir and oblique images improved check point 2D RMSE values from 18 cm to 16 cm and from 54 cm to 25 cm for two different test areas. Our investigations without GCPs delivered similar horizontal values for the dataset Basel15, i.e. accuracy increase from 14 cm to 11 cm and from 51 cm to 34 cm (see section 4.2.4). In case of the dataset Zug17, we improved a 2D RMSE value of 9 cm from direct georeferencing to a value of 5-7 cm for different configurations (see section 4.3.3), which is more accurate by a factor of ca. 3 compared to Fanta-Jende et al. (2019). Same as in our scenarios with no GCPs, they struggle to improve absolute accuracies of the height component. However, by only employing one 3D reference point, our approach enables to achieve absolute 3D accuracies within the sub-decimeter range, which is a typical requirement in 3D mapping projects. Moreover, by incorporating a few more GCPs, we are able to consistently obtain 3D accuracies at the centimeter level.

Compared to Fanta-Jende et al. (2019), we are not dependent on the availability and on the quality of additional derived products. Typical GSDs of ca. 10 cm for nadir and oblique airborne images limit the accuracy potential. Furthermore, possible inaccuracies of aerial image orientations are propagated to the mobile mapping trajectories. While occlusions of road markings by vehicles or vegetation have a significant impact on the performance of the nadir procedure, the oblique approach presumes visible multi-story building façades. In contrast, our integrated georeferencing approach is not limited to urban road environments, but also successfully copes with suburban and rural areas as well as indoor environments. Moreover, our procedure supports varying multi-camera configurations and camera models. We performed all our extensive investigations using real-world scenarios, and did not artificially distort any trajectories.

Recommendations for Multi-Camera Configurations

Multi-camera mobile mapping systems allow for efficient data capturing of road and rail corridors as well as indoor environments. Depending on the main objective of a project, adequate camera types and camera head numbers may vary. While panorama cameras are suitable for image collection in nearly all directions at the same point of time, stereo pinhole camera systems featuring physical bases enable high accuracies. If feasible, a combination of these two types is the prime choice, leading to high image redundancies in multiple directions. Leveraging this redundancy is an essential part for obtaining accurate and robust image orientations, as well as accurate, reliable and dense 3D scene information. However, dense image matching for accurate 3D geometry computation highly relies on adequate stereo bases. Hence, a stereo camera system pointing forward featuring the highest geometric resolution is useful, as the entire road or rail corridor can be mapped. Physical stereo bases in the other directions are not necessarily required, since cameras that do not point in the driving direction can deliver stereo images based on two epochs. However, such virtual stereo bases are entirely dependent on the georeferencing quality and might be less precise. For integrated georeferencing, it is more crucial to connect multi-view imagery from different trajectories than having stereo images all around. Therefore, suitable configurations for efficient data processing complement a stereo camera system pointing forward with mono cameras directed back-right

and back-left or just a camera looking backward. Mapping in opposite driving directions leads to more homogeneous integrated georeferencing solutions and especially stabilizes the height component.

6.2 Limitations and Future Work

Our integrated georeferencing approach is able to deliver accurate and robust image orientations for small to medium-sized scenarios. Processing of up to ca. 10'000 images on a laptop still leads to reasonable computation times. A hierarchical image orientation approach would allow for a speed-up, but highly accurate sub-model alignment is challenging. Hence, parallel processing on scalable computing environments and exploiting high-quality graphics processing units (GPUs) is required for handling large-scale multi-camera datasets.

We rely on a conventional SfM procedure that exploits SIFT features, which works well for our challenging use cases. Nonetheless, we could easily replace the modules of feature extraction and matching with alternative learning-based approaches, that would presumably lead to even better results. Furthermore, such methods could help to reduce our long computation times, in particular for feature matching. Moreover, especially rail and indoor datasets would benefit from a better feature distribution in individual images, which can be obtained by patterns or semantics. Consequently, promising learning-based approaches are reviewed in the following sections.

Both Balntas et al. (2017) and Schönberger et al. (2017) evaluate hand-crafted as well as learned local features. Balntas et al. (2017) show that a simple normalization of traditional hand-crafted descriptors can boost their performance to the level of deep learning-based descriptors. Besides descriptor matching, Schönberger et al. (2017) also evaluate camera poses obtained by SfM based on different feature descriptors. They conclude that advanced hand-crafted features still perform on par or better than learned features in the practical context of image-based reconstruction. Meanwhile, more powerful learned local features have been introduced, e.g. SuperPoint (DeTone et al., 2018), LF-Net (Ono et al., 2018), D2-Net (Dusmanu et al., 2019), R2D2 (Revaud et al., 2019), SOSNet (Tian et al., 2019). These and many others are evaluated in a comprehensive benchmark for local features and robust estimation algorithms, which was recently introduced by Jin et al. (2020). Similar to Schönberger et al. (2017), the accuracy of reconstructed camera poses serves as primary metric, but they differentiate the tasks of wide-baseline stereo and multi-view reconstruction (SfM). Jin et al. (2020) show that classical solutions may still outperform the perceived state of the art with proper settings. Furthermore, they state that end-to-end learning solutions do not yet outperform classical methods that subdivide the problem into separate steps. Hence, Sarlin et al. (2020) propose to learn feature matching with graph neural networks and Kluger et al. (2020) introduce CONSAC, which is the first learning-based method for robust multi-model fitting. Dusmanu et al. (2020) jointly optimize keypoint locations over multiple views according to a non-linear least squares formulation. In addition to improving poor keypoint localization of recent learned feature approaches, they can even refine SIFT features. Wei et al. (2020) present a well performing SfM deep learning framework, which explicitly enforces photometric and geometric consistency as well as camera motion constraints.

There has been a significant amount of work related to visual localization in recent years. While it aims at estimating 6 DoF camera poses such as SfM, it assumes the scene structure to be known. Sattler et al. (2018) introduce a benchmark that motivates the development of robust visual localization approaches. These should be able to handle large appearance variations caused by changes in seasonal and illumination conditions. Kendall et al. (2015) proposed the first end-to-end visual localization approach based on convolutional neural networks (CNNs). These methods learn to directly regress the camera pose from an input image, but according to Sattler et al. (2019), they do not achieve the same level of pose accuracy as conventional 3D structure-based methods. A reason is that pose regression is more closely related to pose approximation via image retrieval than to accurate pose estimation via 3D geometry. Furthermore, there is no guarantee that absolute pose regression approaches generalize beyond their training data. Since structure-based methods rely on the laws of projective geometry as well as on the underlying 3D geometry of the scene, they can better handle viewpoint changes (Sattler et al., 2019). Aiming to deploy visual localization at global-scale, Lynen et al. (2020) do not employ learning-based approaches, but methods using local features and sparse 3D models. This allows for low-latency localization queries and efficient

fusion that is run in real-time on mobile platforms by combining server-side localization with real-time visual-inertial-based camera pose tracking.

Several researchers replace parts of the visual localization pipeline with learning-based components. Meyer et al. (2020) address robust long-term visual localization in large-scale urban environments exploiting street-level imagery. They first perform 2D image-based localization using image retrieval by NetVLAD (Arandjelovic et al., 2016) in order to select corresponding reference images. Then, same as Widya et al. (2018), dense CNN features are extracted on a regular grid, followed by keypoint relocalization for accuracy improvement. Subsequently, feature correspondences are imported in COLMAP for camera pose estimation. Sarlin et al. (2019) propose a robust hierarchical localization pipeline based on the CNN architectures NetVLAD (Arandjelovic et al., 2016) and SuperPoint (DeTone et al., 2018). Furthermore, they introduce HF-Net, a novel CNN that computes keypoints as well as global and local descriptors in a single shot, which leads to reduced computation times.

Schönberger et al. (2018) propose a novel approach to visual localization based on 3D geometric and semantic information, which is able to correctly localize images with strong viewpoint, illumination and seasonal changes. Their model learns a general semantic scene representation in a self-supervised fashion from data, eliminating the need for hand-crafted solutions or manual labeling. Taira et al. (2019) present a pose verification approach to improve large-scale indoor camera localization. In order to cope with repetitive structures and weakly textured scenes, they successfully combine different modalities, namely visual appearance, surface normals and semantics.

Learning-based approaches are meanwhile not only ubiquitous for image orientation, but also for 3D scene generation. While many methods learn individual components of the whole pipeline, end-to-end procedures exist as well. Janai et al. (2020) provide a recent survey of deep learning approaches for both stereo matching and multi-view stereo from an autonomous driving perspective. Zbontar and LeCun (2015) were the first to employ a convolutional neural network for the stereo matching problem. Seki and Pollefeys (2017) and Schönberger et al. (2018) perform SGM using learning-based methods, while Huang et al. (2018) utilize a deep CNN for multi-view stereo reconstruction.

Several of these recent learning-based developments can benefit from integrated georeferencing, but in particular visual localization. In order to reach sufficient accuracy and robustness in varying conditions, diverse georeferenced imagery is needed for the training process. Furthermore, query images rely on existing scene information, thus requiring large-scale image data bases that are accurately georeferenced. Hence, our integrated georeferencing approach is an excellent choice for establishing such high-quality reference data, showing very good performance in both outdoor and indoor environments.

Chapter 7

Appendix

7.1 Camera Pose Computation

Camera pose estimation aims at computing position and attitude information for all images at capturing time. This frequently includes six exterior orientation parameters, i.e. 3D coordinates of projection center in a predefined coordinate reference frame and three Euler angles. However, 3D attitude can also be defined by a rotation matrix or a quaternion. Please note that the computation process for Euler angles conversion, Euler angles to quaternion, and quaternion to Euler angles in section 7.1.2 is indicated sequentially, i.e. a variable of a particular equation refers to the result of the previous computation step. The content of the following sections is derived from multiple sources, e.g. Colomina and Parés (2012).

7.1.1 Coordinate Reference Frames

A coordinate reference frame, often termed frame, consists of both a reference frame that is the realization of a reference system and a coordinate system, which is the parametrization. We briefly introduce the most important ones that are needed in order to understand our computations.

Mapping Frame is the frame in which resulting coordinates are typically demanded. It includes a global terrestrial reference frame as well as horizontal map-projected coordinates and vertical heights. In case of Switzerland, the horizontal reference frames LV03 and LV95 are primarily used. The Swiss vertical reference frame LN02 features leveled heights, whereas LHN95 is based on orthometric heights.

Body Frame is often used for navigation purposes. Its origin is physically located in the navigation center of the IMU. The IMU defines the axes of the body frame: the x-axis points forward, the y-axis points to the right, and the z-axis points downwards.

Camera Frame is identical to the respective two-dimensional image coordinate system, but features a third axis (z) that is the camera axis. The x-axis points to the right, the y-axis points upwards, and the z-axis complements the right-handed coordinate system. The projection center serves as origin.

7.1.2 Rotation Parametrization

We use rotation matrices $\mathbf{R}$, quaternions q_w, q_x, q_y, q_z and Euler angles represented as roll γ, pitch θ, heading ψ or omega ω, phi φ, kappa κ for 3D attitude description. Rotation matrices $\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$ uniquely determine 3D attitude and they are orthogonal with the property that the inverse matrix equals

the transpose: $\mathbf{R}^{-1} = \mathbf{R}^{\mathrm{T}}$ and $\mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{I}$. There are different definitions for rotation matrices, however, we commonly utilize the standard rotation matrix in photogrammetry featuring rotations about the moving axes of the source system (equation 2.27 in Luhmann et al. (2014)):

$$\mathbf{R}_{PG} = \mathbf{R}_{\omega}\mathbf{R}_{\varphi}\mathbf{R}_{\kappa} = \begin{bmatrix} \cos\varphi\cos\kappa & -\cos\varphi\sin\kappa & \sin\varphi \\ \cos\omega\sin\kappa + \sin\omega\sin\varphi\cos\kappa & \cos\omega\cos\kappa - \sin\omega\sin\varphi\sin\kappa & -\sin\omega\cos\varphi \\ \sin\omega\sin\kappa - \cos\omega\sin\varphi\cos\kappa & \sin\omega\cos\kappa + \cos\omega\sin\varphi\sin\kappa & \cos\omega\cos\varphi \end{bmatrix} \tag{7.1}$$

This rotation matrix specifies a projection from image to object coordinates $\mathbf{R}_{PG-img-to-obj}$ with regard to a camera frame mentioned in section 7.1.1. In contrast, the computer vision community frequently uses rotation matrices that specify a projection from object to image coordinates $\mathbf{R}_{CV-obj-to-img}$. Moreover, the x-axis of the underlying coordinate system points to the right, the y-axis points downwards, and the z-axis points away. The relation between these two rotation matrices is given as follows:

$$\mathbf{R}_{CV} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \mathbf{R}_{PG}^{\mathrm{T}} \tag{7.2}$$

Instead of using ambiguous trigonometric functions as in equation (7.1), rotation matrices with algebraic functions can be utilized. Hence, the three independent rotations are described by four algebraic parameters, called quaternions. A quaternion, $\mathbf{q} \in \mathbb{H}$, may be represented as a vector $\mathbf{q} = \begin{bmatrix} q_w & q_x & q_y & q_z \end{bmatrix}^{\mathrm{T}}$. According to Luhmann et al. (2014), a rotation matrix with algebraic functions offers the following benefits:

- no use of trigonometric functions
- simplified computation of the design matrix and faster convergence in adjustment systems
- no singularities
- faster computation by avoiding power series for internal trigonometric calculations

Euler Angles Conversion

From Euler angles representation roll γ, pitch θ, heading ψ to Euler angles representation omega ω, phi φ, kappa κ:

$$\mathbf{R}_{\gamma\theta\psi} = \begin{bmatrix} cos\theta\cos\psi & -\cos\gamma\sin\psi + \sin\gamma\sin\theta\cos\psi & \sin\gamma\sin\psi + \cos\gamma\sin\theta\cos\psi \\ \cos\theta\sin\psi & \cos\gamma\cos\psi + \sin\gamma\sin\theta\sin\psi & -\sin\gamma\cos\psi + \cos\gamma\sin\theta\sin\psi \\ -\sin\theta & \sin\gamma\cos\theta & \cos\gamma\cos\theta \end{bmatrix} \tag{7.3}$$

$$\mathbf{R}_{PG} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \mathbf{R}_{\gamma\theta\psi} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{7.4}$$

$$\varphi = \arctan\left(\frac{r_{13}}{\sqrt{r_{23}^2 + r_{33}^2}}\right) \tag{7.5}$$

- if $\cos\varphi \neq 0$ then $\omega = \arctan\left(-\frac{r_{23}}{r_{33}}\right)$ and $\kappa = \arctan\left(-\frac{r_{12}}{r_{11}}\right)$
- if $\cos\varphi = 0$ (singularity) then $\omega = 0$ and $\kappa = \arctan\left(\frac{r_{22}}{r_{32}}\right)$, however, if $\varphi \neq \frac{\pi}{2}$ then $\kappa = -\arctan\left(\frac{r_{22}}{r_{32}}\right)$

Euler Angles to Quaternion

From Euler angles representation omega ω, phi φ, kappa κ to quaternion q_w, q_x, q_y, q_z:

$$\mathbf{R}_{PG} = \mathbf{R}_{\omega}\mathbf{R}_{\varphi}\mathbf{R}_{\kappa} = \begin{bmatrix} \cos\varphi\cos\kappa & -\cos\varphi\sin\kappa & \sin\varphi \\ \cos\omega\sin\kappa + \sin\omega\sin\varphi\cos\kappa & \cos\omega\cos\kappa - \sin\omega\sin\varphi\sin\kappa & -\sin\omega\cos\varphi \\ \sin\omega\sin\kappa - \cos\omega\sin\varphi\cos\kappa & \sin\omega\cos\kappa + \cos\omega\sin\varphi\sin\kappa & \cos\omega\cos\varphi \end{bmatrix} \tag{7.6}$$

$$\mathbf{R}_{CV} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \mathbf{R}_{PG}^{\mathrm{T}} \tag{7.7}$$

$$\mathbf{K}_1 = \begin{bmatrix} r_{11} - r_{22} - r_{33} & 0 & 0 & 0 \\ r_{12} + r_{21} & r_{22} - r_{11} - r_{33} & 0 & 0 \\ r_{13} + r_{31} & r_{23} + r_{32} & r_{33} - r_{11} - r_{22} & 0 \\ r_{32} - r_{23} & r_{13} - r_{31} & r_{21} - r_{12} & r_{11} + r_{22} + r_{33} \end{bmatrix} \tag{7.8}$$

$$\mathbf{K}_2 = \frac{\mathbf{K}_1}{3} \tag{7.9}$$

- Quaternion q_w, q_x, q_y, q_z is the eigenvector of the symmetric matrix $\mathbf{K}_2$ that corresponds to the largest eigenvalue

Quaternion to Euler Angles

From quaternion q_w, q_x, q_y, q_z to Euler angles representation omega ω, phi φ, kappa κ:

- Quaternion normalization, then

$$\mathbf{R}_{CV} = \begin{bmatrix} 1 - 2q_y^2 - 2q_z^2 & 2q_xq_y - 2q_wq_z & 2q_xq_z + 2q_wq_y \\ 2q_xq_y + 2q_wq_z & 1 - 2q_x^2 - 2q_z^2 & 2q_yq_z - 2q_wq_x \\ 2q_xq_z - 2q_wq_y & 2q_yq_z + 2q_wq_x & 1 - 2q_x^2 - 2q_y^2 \end{bmatrix} \tag{7.10}$$

$$\mathbf{R}_{PG} = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \mathbf{R}_{CV} \right)^{\mathrm{T}} \tag{7.11}$$

$$\varphi = \arctan\left(\frac{r_{13}}{\sqrt{r_{23}^2 + r_{33}^2}} \right) \tag{7.12}$$

- if $\cos\varphi \neq 0$ then $\omega = \arctan\left(-\frac{r_{23}}{r_{33}}\right)$ and $\kappa = \arctan\left(-\frac{r_{12}}{r_{11}}\right)$
- if $\cos\varphi = 0$ (singularity) then $\omega = 0$ and $\kappa = \arctan\left(\frac{r_{22}}{r_{32}}\right)$, however, if $\varphi \neq \frac{\pi}{2}$ then $\kappa = -\arctan\left(\frac{r_{22}}{r_{32}}\right)$

7.2 Publications

The publications created during the course of this thesis are grouped by topic and sorted by date.

Image-Based and Integrated Georeferencing

Cavegn, S., S. Blaser, S. Nebiker, and N. Haala (2018). Robust and Accurate Image-Based Georeferencing Exploiting Relative Orientation Constraints. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume IV-2, Riva del Garda, Italy, pp. 57–64.

Cavegn, S., S. Nebiker, and N. Haala (2016). A Systematic Comparison of Direct and Image-Based Georeferencing in Challenging Urban Areas. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XLI-B1, Prague, Czech Republic, pp. 529–536.

Cavegn, S., S. Nebiker, and N. Haala (2016). Ein systematischer Vergleich zwischen direkter und bildbasierter Georeferenzierung von Mobile Mapping-Stereosequenzen in einem anspruchsvollen Stadtgebiet. In *DGPF Tagungsband 25 / 2016*, Bern, Switzerland, pp. 113–123.

In-Sequence Dense Image Matching

Cavegn, S. and N. Haala (2016). Image-Based Mobile Mapping for 3D Urban Data Capture. *Photogrammetric Engineering & Remote Sensing 82* (12), 925–933.

Nebiker, S., S. Cavegn, and B. Loesch (2015). Cloud-Based Geospatial 3D Image Spaces – A Powerful Urban Model for the Smart City. *ISPRS International Journal of Geo-Information 4* (4), 2267–2291.

Cavegn, S., N. Haala, S. Nebiker, M. Rothermel, and T. Zwölfer (2015). Evaluation of Matching Strategies for Image-Based Mobile Mapping. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume II-3/W5, La Grande Motte, France, pp. 361–368.

Dense Image Matching for Oblique Aerial Scenarios

Schär, P., S. Cavegn, D. Novak, B. Loesch, H. Eugster, and S. Nebiker (2018). Ein systematischer Vergleich verschiedener Multi-View Stereo-Lösungen fur die luftbildgestützte dreidimensionale Infrastrukturkartierung. In *DGPF Tagungsband 27 / 2018*, Munich, Germany, pp. 431–449.

Haala, N. and S. Cavegn (2016). High Density Aerial Image Matching: State-Of-The-Art and Future Prospects. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XLI-B4, Prague, Czech Republic, pp. 625–630.

Haala, N., M. Rothermel, and S. Cavegn (2015). Extracting 3D Urban Models from Oblique Aerial Images. In *Joint Urban Remote Sensing Event (JURSE)*, Lausanne, Switzerland, pp. 1–4.

Haala, N. and S. Cavegn (2015). Benchmark zur Evaluation dichter Bildzuordnungsverfahren in Luftbildern. In *DGPF Tagungsband 24 / 2015*, Cologne, Germany, pp. 244–253.

Cavegn, S., N. Haala, S. Nebiker, M. Rothermel, and P. Tutzauer (2014). Benchmarking High Density Image Matching for Oblique Airborne Imagery. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XL-3, Zurich, Switzerland, pp. 45–52.

Deuber, M., S. Cavegn, and S. Nebiker (2014). Dense Image Matching. Performance Analysis on Oblique Imagery. *GIM International 28* (9), 23–25.

Cavegn, S., S. Nebiker, and M. Deuber (2014). Dense Image Matching mit Oblique Luftbildaufnahmen – Ein systematischer Vergleich verschiedener Lösungen mit Aufnahmen der Leica RCD30 Oblique Penta. In *DGPF Tagungsband 23 / 2014*, Hamburg, Germany, pp. 1–10.

Mobile Mapping Systems

Blaser, S., S. Cavegn, and S. Nebiker (2018). Development of a Portable High Performance Mobile Mapping System Using the Robot Operating System. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume IV-1, Karlsruhe, Germany, pp. 13–20.

Blaser, S., S. Nebiker, and S. Cavegn (2018). On a Novel 360° Panoramic Stereo Mobile Mapping System. *Photogrammetric Engineering & Remote Sensing 84* (6), 347–356.

Blaser, S., S. Nebiker, and S. Cavegn (2017). System Design, Calibration and Performance Analysis of a Novel 360° Stereo Panoramic Mobile Mapping System. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume IV-1/W1, Hannover, Germany, pp. 207–213.

Visual Localization

Rettenmund, D., M. Fehr, S. Cavegn, and S. Nebiker (2018). Accurate Visual Localization in Outdoor and Indoor Environments Exploiting 3D Image Spaces as Spatial Reference. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XLII-1, Karlsruhe, Germany, pp. 355–362.

Bibliography

Abraham, S. and W. Förstner (2005). Fish-eye-stereo calibration and epipolar rectification. *ISPRS Journal of Photogrammetry and Remote Sensing 59*(5), 278–288.

Ackermann, F., H. Ebner, and H. Klein (1970). Ein Rechenprogramm für die Aerotriangulation mit unabhängigen Modellen. *Bildmessung und Luftbildwesen 4*, 206–217.

Agarwal, S., K. Mierle, and Others (2020). Ceres Solver. http://ceres-solver.org.

Agarwal, S., N. Snavely, S. M. Seitz, and R. Szeliski (2010). Bundle Adjustment in the Large. In K. Daniilidis, P. Maragos, and N. Paragios (Eds.), *ECCV 2010, Part II, LNCS 6312*, Heraklion, Greece, pp. 29–42. Springer Berlin Heidelberg.

Ahmadabadian, A. H., S. Robson, J. Boehm, M. Shortis, K. Wenzel, and D. Fritsch (2013). A comparison of dense matching algorithms for scaled surface reconstruction using stereo camera rigs. *ISPRS Journal of Photogrammetry and Remote Sensing 78*(2013), 157–167.

Anguelov, D., C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver (2010). Google Street View: Capturing the World at Street Level. *IEEE Computer 43*(6), 32–38.

Arandjelovic, R., P. Gronat, A. Torii, T. Pajdla, and J. Sivic (2016). NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 5297–5307.

Balntas, V., K. Lenc, A. Vedaldi, and K. Mikolajczyk (2017). HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 5173–5182.

Barber, D., J. Mills, and S. Smith-Voysey (2008). Geometric validation of a ground-based mobile laser scanning system. *ISPRS Journal of Photogrammetry & Remote Sensing 63*(1), 128–141.

Bayoud, F. A. (2006). *Development of a Robotic Mobile Mapping System by Vision-Aided Inertial Navigation: A Geomatics Approach*. Phd, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

Berger, M., A. Tagliasacchi, L. Seversky, P. Alliez, J. Levine, A. Sharf, and C. Silva (2014). State of the Art in Surface Reconstruction from Point Clouds. In *Eurographics 2014 - State of the Art Reports*, Strasbourg, France, pp. 161–185.

Bethmann, F. and T. Luhmann (2017). Object-Based Semi-global Multi-image Matching. *PFG - Journal of Photogrammetry, Remote Sensing and Geoinformation Science 85*(6), 349–364.

Blaser, S., S. Cavegn, and S. Nebiker (2018). Development of a Portable High Performance Mobile Mapping System Using the Robot Operating System. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume IV-1, Karlsruhe, Germany, pp. 13–20.

Blaser, S., S. Nebiker, and S. Cavegn (2018). On a Novel 360 Panoramic Stereo Mobile Mapping System. *Photogrammetric Engineering & Remote Sensing 84*(6), 347–356.

Brenner, C. and S. Hofmann (2012). Evaluation of Automatically Extracted Landmarks for Future Driver Assistance Systems. In A. Yeh, W. Shi, Y. Leung, and C. Zhou (Eds.), *Advances in Spatial Data Handling and GIS. Lecture Notes in Geoinformation and Cartography*, pp. 169–181. Springer Berlin Heidelberg.

Burkhard, J., S. Cavegn, A. Barmettler, and S. Nebiker (2012). Stereovision Mobile Mapping: System Design and Performance Evaluation. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XXXIX-B5, Melbourne, Australia, pp. 453–458.

Cadena, C., L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard (2016). Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Transactions on Robotics 32*(6), 1309–1332.

Cavegn, S. and N. Haala (2016). Image-Based Mobile Mapping for 3D Urban Data Capture. *Photogrammetric Engineering & Remote Sensing 82*(12), 925–933.

Cavegn, S., N. Haala, S. Nebiker, M. Rothermel, and P. Tutzauer (2014). Benchmarking High Density Image Matching for Oblique Airborne Imagery. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XL-3, Zurich, Switzerland, pp. 45–52.

Cavegn, S., N. Haala, S. Nebiker, M. Rothermel, and T. Zwölfer (2015). Evaluation of Matching Strategies for Image-Based Mobile Mapping. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume II-3/W5, La Grande Motte, France, pp. 361–368.

Cefalu, A. and D. Fritsch (2014). Non-Incremental Derivation of Scale and Pose from a Network of Relative Orientations. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XL-3, Zurich, Switzerland, pp. 53–59.

Cefalu, A., N. Haala, and D. Fritsch (2016). Structureless Bundle Adjustment With Self-Calibration Using Accumulated Constraints. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume III-3, Prague, Czech Republic, pp. 3–9.

Collins, R. T. (1996). A Space-Sweep Approach to True Multi-Image Matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, pp. 358–363.

Colomina, I. and P. Molina (2014). Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing 92*, 79–97.

Colomina, I. and M. E. Parés (2012). Sensor Orientation: Precise Trajectory and Attitude Determination with INS. In *Calibration and Orientation Workshop (EuroCOW)*, Castelldefels, Spain, pp. 457.

Cramer, M. (2001). *Genauigkeitsuntersuchungen zur GPS/INS-Integration in der Aerophotogrammetrie*. Phd, University of Stuttgart, Germany.

Cucci, D. A., M. Rehak, and J. Skaloud (2017). Bundle adjustment with raw inertial observations in UAV applications. *ISPRS Journal of Photogrammetry and Remote Sensing 130*, 1–12.

Cui, H., X. Gao, S. Shen, and Z. Hu (2017). HSfM: Hybrid Structure-from-Motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 1212–1221.

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

Cui, Z. and P. Tan (2015). Global Structure-from-Motion by Similarity Averaging. In *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 864–872.

Dellaert, F. (2012). Factor Graphs and GTSAM: A Hands-on Introduction. Technical report.

DeTone, D., T. Malisiewicz, and A. Rabinovich (2018). SuperPoint: Self-Supervised Interest Point Detection and Description. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, pp. 337–349.

Dong, J. and S. Soatto (2015). Domain-Size Pooling in Local Descriptors: DSP-SIFT. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, pp. 5097–5106.

Dusmanu, M., I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler (2019). D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, pp. 8092–8101.

Dusmanu, M., J. L. Schönberger, and M. Pollefeys (2020). Multi-View Optimization of Local Feature Geometry. In *ECCV 2020*, pp. 1–16.

Ellum, C. and N. El-Sheimy (2006). New Strategies for Integrating Photogrammetric and GNSS Data. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XXXVI-5, Dresden, Germany, pp. 103–108.

Eugster, H. (2011). *Echtzeit-Georegistrierung von Videodaten mit Hilfe von Navigationssensoren geringer Qualität und digitalen 3D-Lanschaftsmodellen.* Phd, Humboldt-Universität zu Berlin, Germany.

Eugster, H., F. Huber, S. Nebiker, and A. Gisi (2012). Integrated Georeferencing of Stereo Image Sequences Captured with a Stereovision Mobile Mapping System - Approaches and Practical Results. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XXXIX-B1, Melbourne, Australia, pp. 309–314.

Fanta-Jende, P., F. Nex, M. Gerke, J. Lijnen, and G. Vosselman (2019). Correction of Mobile Mapping Trajectories in GNSS-Denied Environments Using Aerial Nadir and Aerial Oblique Images. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XLII-2/W13, Enschede, The Netherlands, pp. 1649–1654.

Fanta-Jende, P., F. Nex, G. Vosselman, and M. Gerke (2019). Co-registration of panoramic mobile mapping images and oblique aerial images. *The Photogrammetric Record 34*(166), 148–173.

Fischler, M. a. and R. C. Bolles (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM 24*(6), 381–395.

Forlani, G., R. Roncella, and F. Remondino (2005). Structure and Motion Reconstruction of Short Mobile Mapping Image Sequences. In *VII Conference on Optical 3D Measurement Techniques*, Vienna, Austria, pp. 265–274.

Forster, C., L. Carlone, F. Dellaert, and D. Scaramuzza (2016). On-Manifold Preintegration for Real-Time Visual-Inertial Odometry. *IEEE Transactions on Robotics*, 1–21.

Frahm, J.-M., M. Pollefeys, S. Lazebnik, D. Gallup, B. Clipp, R. Raguram, C. Wu, C. Zach, and T. Johnson (2010). Fast robust large-scale mapping from video and internet photo collections. *ISPRS Journal of Photogrammetry and Remote Sensing 65*(6), 538–549.

Fraundorfer, F. and D. Scaramuzza (2012). Visual Odometry, Part II: Matching, Robustness, Optimization, and Applications. *IEEE Robotics & Automation Magazine 19*(2), 78–90.

Furukawa, Y. and J. Ponce (2010). Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*(8), 1362–1376.

Fusiello, A., E. Trucco, and A. Verri (2000). A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications 12*(1), 16–22.

Gallup, D. (2011). *Efficient 3D Reconstruction of Large-Scale Urban Environments from Street-Level Video.* Phd, University of North Carolina, NC.

Geiger, A., P. Lenz, and R. Urtasun (2012). Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, pp. 3354–3361.

Geppert, M., P. Liu, Z. Cui, M. Pollefeys, and T. Sattler (2019). Efficient 2D-3D Matching for Multi-Camera Visual Localization. In *IEEE International Conference on Robotics and Automation (ICRA)*, Montreal, Canada, pp. 5972–5978.

Gerke, M. (2009). Dense Matching in High Resolution Oblique Airborne Images. In U. Stilla, F. Rottensteiner, and N. Paparoditis (Eds.), *CMRT09, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XXXVIII-3, Paris, France, pp. 77–82.

Gherardi, R., M. Farenzena, and A. Fusiello (2010). Improving the Efficiency of Hierarchical Structure-and-Motion. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, pp. 1594–1600.

Grün, A. (1982). The Accuracy Potential of the Modern Bundle Block Adjustment in Aerial Photogrammetry. *Photogrammetric Engineering and Remote Sensing 48*(1), 45–54.

Haala, N. (2005). *Multi-Sensor-Photogrammetrie.* Habilitationsschrift, University of Stuttgart, Germany.

Haala, N. (2014). Benchmark on Image Matching. Final Report. EuroSDR-Project, Commission 2. EuroSDR Publication Series, Official Publication No. 64. Technical report.

Haala, N., M. Peter, J. Kremer, and G. Hunter (2008). Mobile Lidar Mapping for 3D Point Cloud Collection in Urban Areas - A Performance Test. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XXXVII-B5, Beijing, China, pp. 1119–1124.

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

Hassan, T., C. Ellum, and N. El-Sheimy (2006). Bridging Land-Based Mobile Mapping Using Photogrammetric Adjustments. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XXXVI-1, Paris, France.

Havlena, M., T. Pajdla, and K. Cornelis (2008). Structure from Omnidirectional Stereo Rig Motion for City Modeling. In *VISAPP08*, pp. 407–414.

Havlena, M., A. Torii, and T. Pajdla (2010). Efficient Structure from Motion by Graph Optimization. In K. Daniilidis, P. Maragos, and N. Paragios (Eds.), *ECCV 2010, Part II, LNCS 6312*, Heraklion, Greece, pp. 100–113. Springer Berlin Heidelberg.

Heipke, C., K. Jacobsen, and H. Wegmann (2002). Analysis of the Results of the OEEPE Test "Integrated Sensor Orientation". In C. Heipke, K. Jacobsen, and H. Wegmann (Eds.), *Integrated Sensor Orientation, OEEPE Official Publication No. 43*, pp. 31–49.

Heng, L., B. Choi, Z. Cui, M. Geppert, S. Hu, B. Kuan, P. Liu, R. Nguyen, Y. C. Yeo, A. Geiger, G. H. Lee, M. Pollefeys, and T. Sattler (2019). Project AutoVision: Localization and 3D Scene Perception for an Autonomous Vehicle with a Multi-Camera System. In *IEEE International Conference on Robotics and Automation (ICRA)*, Montreal, Canada, pp. 4695–4702.

Hess, W., D. Kohler, H. Rapp, and D. Andor (2016). Real-Time Loop Closure in 2D LIDAR SLAM. In *IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, pp. 1271–1278.

Hirschmüller, H. (2008). Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*(2), 328–341.

Huang, P.-H., K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang (2018). DeepMVS: Learning Multi-view Stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, pp. 2821–2830.

Hussnain, Z., S. O. Elberink, and G. Vosselman (2019). Automatic extraction of accurate 3D tie points for trajectory adjustment of mobile laser scanners using aerial imagery. *ISPRS Journal of Photogrammetry and Remote Sensing 154*(June), 41–58.

Hussnain, Z., S. Oude Elberink, and G. Vosselman (2018). An Automatic Procedure For Mobile Laser Scanning Platform 6DoF Trajectory Adjustment. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XLII-1, Karlsruhe, Germany, pp. 203–209.

Indelman, V., R. Roberts, and F. Dellaert (2015). Incremental light bundle adjustment for structure from motion and robotics. *Robotics and Autonomous Systems 70*, 63–82.

Irschara, A. (2012). *Scalable Scene Reconstruction and Image Based Localization.* Phd, Graz University of Technology, Austria.

Janai, J., F. Güney, A. Behl, and A. Geiger (2020). Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Foundations and Trends in Computer Graphics and Vision 12*(1-3), 1–308.

Javanmardi, M., E. Javanmardi, Y. Gu, and S. Kamijo (2017). Towards High-Definition 3D Urban Mapping: Road Feature-Based Registration of Mobile Mapping Systems and Aerial Imagery. *Remote Sensing 9*(10), 975.

Jende, P., F. Nex, M. Gerke, and G. Vosselman (2018). A fully automatic approach to register mobile mapping and airborne imagery to support the correction of platform trajectories in GNSS-denied urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing 141*(July), 86–99.

Jin, Y., D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi, and E. Trulls (2020). Image Matching across Wide Baselines: From Paper to Practice. *International Journal of Computer Vision* (accepted).

Karel, W., C. Ressl, and N. Pfeifer (2016). Efficient Orientation and Calibration of Large Aerial Blocks of Multi-Camera Platforms. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XLI-B1, Prague, Czech Republic, pp. 199–204.

Kendall, A., M. Grimes, and R. Cipolla (2015). PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 2938–2946.

Kersting, A. P., A. Habib, and J.-Y. Rau (2012). New Method for the Calibration of Multi-Camera Mobile Mapping Systems. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XXXIX-B1, Melbourne, Australia, pp. 121–126.

Klingner, B., D. Martin, and J. Roseborough (2013). Street View Motion-from-Structure-from-Motion. In *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, pp. 953–960.

Kluger, F., E. Brachmann, H. Ackermann, C. Rother, M. Y. Yang, and B. Rosenhahn (2020). CONSAC: Robust Multi-Model Fitting by Conditional Sample Consensus. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4634–4643.

Knapitsch, A., J. Park, Q.-Y. Zhou, and V. Koltun (2017). Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics 36*(4), 1–13.

Kneip, L., P. Furgale, and R. Siegwart (2013). Using Multi-Camera Systems in Robotics: Efficient Solutions to the NPnP Problem. In *IEEE International Conference on Robotics and Automation (ICRA)*, Karlsruhe, Germany, pp. 3770–3776.

Kümmerle, R., G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard (2011). g2o: A General Framework for Graph Optimization. In *IEEE International Conference on Robotics and Automation*, Shanghai, China, pp. 3607–3613.

Kuo, J., M. Muglikar, Z. Zhang, and D. Scaramuzza (2020). Redesigning SLAM for Arbitrary Multi-Camera Systems. In *IEEE International Conference on Robotics and Automation (ICRA)*, Paris, France.

Leberl, F., A. Irschara, T. Pock, P. Meixner, M. Gruber, S. Scholz, and A. Wiechert (2010). Point Clouds: Lidar versus 3D Vision. *Photogrammetric Engineering and Remote Sensing 76*(10), 1123–1134.

Liu, P., M. Geppert, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys (2018). Towards Robust Visual Odometry with a Multi-Camera System. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, pp. 1154–1161.

Loop, C. and Z. Zhang (1999). Computing Rectifying Homographies for Stereo Vision. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Fort Collins, CO, pp. 125–131.

Luhmann, T., S. Robson, S. Kyle, and J. Boehm (2014). *Close-Range Photogrammetry and 3D Imaging* (2nd ed.). Berlin, Boston: De Gruyter.

Lynen, S., T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart (2015). Get Out of My Lab: Large-scale, Real-Time Visual-Inertial Localization. In *Robotics: Science and Systems (RSS) XI*, Rome, Italy, pp. 37–46.

Lynen, S., B. Zeisl, D. Aiger, M. Bosse, J. Hesch, M. Pollefeys, R. Siegwart, and T. Sattler (2020). Large-scale, real-time visual-inertial localization revisited. *The International Journal of Robotics Research 39*(9), 1061–1084.

Mapillary (2020). OpenSfM. https://github.com/mapillary/OpenSfM.

Meilland, M., A. I. Comport, and P. Rives (2015). Dense Omnidirectional RGB-D Mapping of Large-scale Outdoor Environments for Real-time Localization and Autonomous Navigation. *Journal of Field Robotics 32*(4), 474–503.

Menze, M. and A. Geiger (2015). Object Scene Flow for Autonomous Vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, pp. 3061–3070.

Meyer, J., D. Rettenmund, and S. Nebiker (2020). Long-Term Visual Localization in Large Scale Urban Environments Exploiting Street Level Imagery. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume V-2-2020, pp. 57–63.

Molina, P., M. Blázquez, D. Cucci, and I. Colomina (2017). First Results of a Tandem Terrestrial-Unmanned Aerial mapKITE System with Kinematic Ground Control Points for Corridor Mapping. *Remote Sensing 9*(1), 13.

Molina, P., M. Blázquez, J. Sastre, and I. Colomina (2016). mapKITE: A New Paradigm for Simultaneous Aerial and Terrestrial Geodata Acquisition and Mapping. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XLI-B1, Prague, Czech Republic, pp. 957–962.

Moulon, P., P. Monasse, R. Perrot, and R. Marlet (2017). OpenMVG: Open Multiple View Geometry. In B. Kerautret, M. Colom, and P. Monasse (Eds.), *International Workshop on Reproducible Research in Pattern Recognition, RRPR 2016, LNCS 10214*, Cancun, Mexico, pp. 60–74. Springer, Cham.

Musialski, P., P. Wonka, D. G. Aliaga, M. Wimmer, L. van Gool, and W. Purgathofer (2013). A Survey of Urban Reconstruction. *Computer Graphics Forum 32*(6), 146–177.

Nebiker, S. (2019). Stereo Image Capturing System. Patent US 2019/0020829 A1.

Nebiker, S., S. Bleisch, and M. Christen (2010). Rich point clouds in virtual globes - A new paradigm in city modeling? *Computers, Environment and Urban Systems 34*(6), 508–517.

Nebiker, S., S. Cavegn, H. Eugster, K. Laemmer, J. Markram, and R. Wagner (2012). Fusion of Airborne and Terrestrial Image-Based 3D Modelling for Road Infrastructure Management - Vision and First Experiments. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XXXIX-B4, Melbourne, Australia, pp. 79–84.

Nebiker, S., S. Cavegn, and B. Loesch (2015). Cloud-Based Geospatial 3D Image Spaces - A Powerful Urban Model for the Smart City. *ISPRS International Journal of Geo-Information 4*(4), 2267–2291.

Nex, F., M. Gerke, F. Remondino, H.-J. Przybilla, M. Bäumker, and A. Zurhorst (2015). ISPRS Benchmark for Multi-Platform Photogrammetry. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume II-3/W4, Munich, Germany, pp. 135–142.

Novak, K. (1991). The Ohio State University Highway Mapping System: The Stereo Vision System Component. In *Proceedings of the 47th Annual Meeting of The Institute of Navigation*, pp. 121–124.

Ono, Y., E. Trulls, P. Fua, and K. M. Yi (2018). LF-Net: Learning Local Features from Images. In *32nd Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, pp. 1–11.

Oram, D. (2001). Rectification for Any Epipolar Geometry. In *Procedings of the British Machine Vision Conference*, pp. 653–662.

Özdemir, E., I. Toschi, and F. Remondino (2019). A Multi-Purpose Benchmark for Photogrammetric Urban 3D Reconstruction in a Controlled Environment. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XLII-1/W2, Warsaw, Poland, pp. 53–60.

Paparoditis, N., J.-P. Papelard, B. Cannelle, A. Devaux, B. Soheilian, N. David, and E. Houzay (2012). Stereopolis II: A Multi-Purpose and Multi-Sensor 3D Mobile Mapping System for Street Visualisation and 3D Metrology. *Revue française de photogrammétrie et de télédétection* (200), 69–79.

Parys, R. and A. Schilling (2012). Incremental Large Scale 3D Reconstruction. In *Second International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, Zurich, Switzerland, pp. 416–423.

Pollefeys, M., R. Koch, and L. Van Gool (1999). A simple and efficient rectification method for general motion. In *IEEE International Conference on Computer Vision (ICCV)*, Kerkyra, Greece, pp. 496–501.

Pollefeys, M., D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles (2008). Detailed Real-Time Urban 3D Reconstruction from Video. *International Journal of Computer Vision 78*(2-3), 143–167.

Puente, I., H. González-Jorge, J. Martínez-Sánchez, and P. Arias (2013). Review of mobile mapping and surveying technologies. *Measurement 46*(7), 2127–2145.

Rehak, M. (2017). *Integrated Sensor Orientation on Micro Aerial Vehicles*. Phd, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

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

Remondino, F., M. G. Spera, E. Nocerino, F. Menna, and F. Nex (2014). State of the Art in High Density Image Matching. *The Photogrammetric Record 29*(146), 144–166.

Revaud, J., P. Weinzaepfel, C. De Souza, and M. Humenberger (2019). R2D2: Repeatable and Reliable Detector and Descriptor. In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, pp. 1–11.

Rodríguez López, A. L. (2013). *Algebraic epipolar constraints for efficient structureless multiview motion estimation*. Phd, Universidad de Murcia, Spain.

Rothermel, M. (2017). *Development of a SGM-Based Multi-View Reconstruction Framework for Aerial Imagery*. Phd, University of Stuttgart, Germany.

Rothermel, M., K. Wenzel, D. Fritsch, and N. Haala (2012). SURE: Photogrammetric Surface Reconstruction from Imagery. In *LC3D Workshop*, Berlin, Germany.

Rumpler, M., A. Tscharf, C. Mostegel, S. Daftry, C. Hoppe, R. Prettenthaler, F. Fraundorfer, G. Mayer, and H. Bischof (2017). Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance. *Computer Vision and Image Understanding 157*, 255–273.

Rupnik, E., M. Daakir, and M. Pierrot Deseilligny (2017). MicMac - a free, open-source solution for photogrammetry. *Open Geospatial Data, Software and Standards 2*(14), 1–9.

Rupnik, E., F. Nex, I. Toschi, and F. Remondino (2015). Aerial multi-camera systems: Accuracy and block triangulation issues. *ISPRS Journal of Photogrammetry and Remote Sensing 101*(60), 233–246.

Sarlin, P.-E., C. Cadena, R. Siegwart, and M. Dymczyk (2019). From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, pp. 12716–12725.

Sarlin, P.-E., D. DeTone, T. Malisiewicz, and A. Rabinovich (2020). SuperGlue: Learning Feature Matching with Graph Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4938–4947.

Sattler, T., W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla (2018). Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, pp. 8601–8610.

Sattler, T., Q. Zhou, M. Pollefeys, and L. Leal-Taixe (2019). Understanding the Limitations of CNN-based Absolute Camera Pose Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, pp. 3302–3312.

Scaramuzza, D. and F. Fraundorfer (2011). Visual Odometry, Part I: The First 30 Years and Fundamentals. *IEEE Robotics & Automation Magazine 18*(4), 80–92.

Schär, P., S. Cavegn, D. Novak, B. Loesch, H. Eugster, and S. Nebiker (2018). Ein systematischer Vergleich verschiedener Multi-View Stereo-Lösungen für die luftbildgestützte dreidimensionale Infrastrukturkartierung. In *DGPF Tagungsband 27 / 2018*, Munich, Germany, pp. 431–449.

Scharstein, D., H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling (2014). High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In X. Jiang, J. Hornegger, and R. Koch (Eds.), *GCPR 2014, LNCS 8753*, Munster, Germany, pp. 31–42. Springer, Cham.

Scharstein, D. and R. Szeliski (2002). A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision 47*(1-3), 7–42.

Schneider, D., E. Schwalbe, and H.-G. Maas (2009). Validation of geometric models for fisheye lenses. *ISPRS Journal of Photogrammetry and Remote Sensing 64*(3), 259–266.

Schneider, J., F. Schindler, T. Läbe, and W. Förstner (2012). Bundle Adjustment for Multi-Camera Systems with Points at Infinity. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume I-3, Melbourne, Australia, pp. 75–80.

Schönberger, J. L. (2018). *Robust Methods for Accurate and Efficient 3D Modeling from Unstructured Imagery.* Phd, ETH Zürich, Switzerland.

Schönberger, J. L. (2020). COLMAP. https://colmap.github.io.

Schönberger, J. L. and J.-M. Frahm (2016). Structure-from-Motion Revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 4104–4113.

Schönberger, J. L., F. Fraundorfer, and J.-M. Frahm (2014). Structure-from-Motion for MAV Image Sequence Analysis with Photogrammetric Applications. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XL-3, Zurich, Switzerland, pp. 305–312.

Schönberger, J. L., H. Hardmeier, T. Sattler, and M. Pollefeys (2017). Comparative Evaluation of Hand-Crafted and Learned Local Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 1482–1491.

Schönberger, J. L., M. Pollefeys, A. Geiger, and T. Sattler (2018). Semantic Visual Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, pp. 6896–6906.

Schönberger, J. L., S. N. Sinha, and M. Pollefeys (2018). Learning to Fuse Proposals from Multiple Scanline Optimizations in Semi-Global Matching. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss (Eds.), *ECCV 2018, LNCS 11217*, Munich, Germany, pp. 758–775. Springer, Cham.

Schönberger, J. L., E. Zheng, M. Pollefeys, and J.-M. Frahm (2016). Pixelwise View Selection for Unstructured Multi-View Stereo. In B. Leibe, J. Matas, N. Sebe, and M. Welling (Eds.), *ECCV 2016, LNCS 9907*, Amsterdam, The Netherlands, pp. 501–518. Springer, Cham.

Schöps, T., J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger (2017). A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 3260–3269.

Schwarz, K., H. Martell, N. El-Sheimy, R. Li, M. Chapman, and D. Cosandier (1993). VIASAT - A mobile highway survey system of high accuracy. In *Proceedings of VNIS '93 - Vehicle Navigation and Information Systems Conference*, pp. 476–481.

Schwarz, K. P., M. A. Chapman, M. E. Cannon, and P. Gong (1993). An Integrated INS/GPS Approach to the Georeferencing of Remotely Sensed Data. *Photogrammetric Engineering & Remote Sensing 59*(11), 1667–1674.

Seitz, S., B. Curless, J. Diebel, D. Scharstein, and R. Szeliski (2006). A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, pp. 519–528.

Seki, A. and M. Pollefeys (2017). SGM-Nets: Semi-global matching with neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 231–240.

Shah, R., A. Deshpande, and P. J. Narayanan (2015). Multistage SFM: A Coarse-to-Fine Approach for 3D Reconstruction. *arXiv:1512.06235*.

Shan, Q., C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz (2014). Accurate Geo-Registration by Ground-to-Aerial Image Matching. In *International Conference on 3D Vision*, Tokyo, Japan, pp. 525–532.

Silva, J. F. C., M. C. Lemes Neto, and V. Blasechi (2014). Automating the Photogrammetric Bridging Based on MMS Image Sequence Processing. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XL-1, Denver, CO, pp. 389–396.

Skaloud, J. (1999). *Optimizing Georeferencing of Airborne Survey Systems by INS/DGPS*. Phd, University of Calgary, Canada.

Steffen, R., J.-M. Frahm, and W. Förstner (2012). Relative Bundle Adjustment Based on Trifocal Constraints. In K. Kutulakos (Ed.), *ECCV 2010 Workshops, Part II, LNCS 6554*, Heraklion, Greece, pp. 282–295. Springer Berlin Heidelberg.

Strecha, C., W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen (2008). On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, pp. 1–8.

Sun, Y., H. Sun, L. Yan, S. Fan, and R. Chen (2016). RBA: Reduced Bundle Adjustment for oblique aerial photogrammetry. *ISPRS Journal of Photogrammetry and Remote Sensing 121*, 128–142.

Sünderhauf, N. (2012). *Robust Optimization for Simultaneous Localization and Mapping*. Phd, Technische Universität Chemnitz, Germany.

Sweeney, C., T. Höllerer, and M. Turk (2015). Theia: A Fast and Scalable Structure-from-Motion Library. In *Proceedings of the 23rd ACM international conference on Multimedia - MM '15*, pp. 693–696. ACM Press.

Szeliski, R. (2011). *Computer Vision. Algorithms and Applications*. London: Springer.

Taira, H., I. Rocco, J. Sedlar, M. Okutomi, J. Sivic, T. Pajdla, T. Sattler, and A. Torii (2019). Is This The Right Place? Geometric-Semantic Pose Verification for Indoor Visual Localization. In *IEEE International Conference on Computer Vision (ICCV)*, Seoul, Korea, pp. 4373–4383.

Tian, Y., X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas (2019). SOSNet: Second Order Similarity Regularization for Local Descriptor Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, pp. 11016–11025.

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

Tournaire, O., B. Soheilian, and N. Paparoditis (2006). Towards a Sub-Decimetric Georeferencing of Ground-Based Mobile Mapping Systems in Urban Areas: Matching Ground-Based and Aerial-Based Imagery Using Roadmarks. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XXXVI-1, Paris, France.

Triggs, B., P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon (2000). Bundle Adjustment - A Modern Synthesis. In B. Triggs, A. Zisserman, and R. Szeliski (Eds.), *Vision Algorithms: Theory and Practice, IWVA 1999, LNCS 1883*, Corfu, Greece, pp. 298–372. Springer Berlin Heidelberg.

Tron, R., X. Zhou, and K. Daniilidis (2016). A Survey on Rotation Optimization in Structure from Motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Las Vegas, NV, pp. 1032–1040.

Urban, S., S. Wursthorn, J. Leitloff, and S. Hinz (2017). MultiCol Bundle Adjustment: A Generic Method for Pose Estimation, Simultaneous Self-Calibration and Reconstruction for Arbitrary Multi-Camera Systems. *International Journal of Computer Vision 121*(2), 234–252.

Van Den Heuvel, F. A., R. Verwaal, and B. Beers (2006). Calibration of Fisheye Camera Systems and the Reduction of Chromatic Aberration. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XXXVI-5, Dresden, Germany.

Wei, X., Y. Zhang, Z. Li, Y. Fu, and X. Xue (2020). DeepSFM: Structure From Motion Via Deep Bundle Adjustment. In *ECCV 2020*, pp. 1–17.

Wenzel, K. (2016). *Dense Image Matching for Close Range Photogrammetry*. Phd, University of Stuttgart, Germany.

Wenzel, K., M. Rothermel, D. Fritsch, and N. Haala (2014). Filtering of Point Clouds from Photogrammetric Surface Reconstruction. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Volume XL-5, Riva del Garda, Italy, pp. 615–620.

Widya, A. R., A. Torii, and M. Okutomi (2018). Structure from motion using dense CNN features with keypoint relocalization. *IPSJ Transactions on Computer Vision and Applications 10*(6), 1–7.

Wu, B., L. Xie, H. Hu, Q. Zhu, and E. Yau (2018). Integration of aerial oblique imagery and terrestrial imagery for optimized 3D modeling in urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing 139*, 119–132.

Wu, C. (2013). Towards Linear-Time Incremental Structure from Motion. In *International Conference on 3D Vision*, Seattle, WA, pp. 127–134.

Yousif, K., A. Bab-Hadiashar, and R. Hoseinnezhad (2015). An Overview to Visual Odometry and Visual SLAM: Applications to Mobile Robotics. *Intelligent Industrial Systems 1*(4), 289–311.

Zbontar, J. and Y. LeCun (2015). Computing the Stereo Matching Cost with a Convolutional Neural Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, pp. 1592–1599.

Zhao, L., S. Huang, and G. Dissanayake (2018). Linear SFM: A hierarchical approach to solving structure-from-motion problems by decoupling the linear and nonlinear components. *ISPRS Journal of Photogrammetry and Remote Sensing 141*, 275–289.

Zheng, E. and C. Wu (2015). Structure from Motion Using Structure-Less Resection. In *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 2075–2083.

Zhu, S., R. Zhang, L. Zhou, T. Shen, T. Fang, P. Tan, and L. Quan (2018). Very Large-Scale Global SfM by Distributed Motion Averaging. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, pp. 4568–4577.

Acknowledgements

Numerous people were involved in the creation of this dissertation. First of all, I would like to express my deep gratitude and appreciation to my thesis supervisors apl. Prof. Dr.-Ing. Norbert Haala and Prof. Dr. Stephan Nebiker for their continuous guidance, support, encouragement and trust. Moreover, I would like to thank Prof. Dr.-Ing. habil. Dr. h.c. Volker Schwieger for agreeing to be a co-referee and for the valuable feedback on this thesis. I greatly benefited from being associated with two research institutes, namely the Institute for Photogrammetry (ifp) at the University of Stuttgart as well as the Institute of Geomatics (IGEO) at the University of Applied Sciences and Arts Northwestern Switzerland (FHNW). While the ifp gave me the opportunity to conduct a PhD, my employment at the IGEO facilitated to perform interesting and challenging research work. I am very grateful to all my colleagues at both institutions who allowed for a stimulating work atmosphere and supported me to succeed. During my employment at the IGEO, I had the opportunity to work on three mobile mapping projects in collaboration with the company iNovitas, i.e. infraVIS (Sustainable Infrastructure Management based on Versatile Intelligent 3D Image Spaces), BIMAGE (Building Information Management based on Geospatial 3D Imagery) and cloudIO (Cloud-based Image Orientation for Infrastructure Management). Hence, I am not only thankful to the FHNW and iNovitas, but also to the Swiss Innovation Agency (Innosuisse) who mainly funded these research projects. Furthermore, I would like to thank iNovitas and the respective agencies for providing the vehicle-based mobile mapping datasets Basel14, Zug17 and Vienna16. Thanks are also due to the company nFrames for providing code and licenses of the SURE software system that enables photogrammetric 3D surface reconstruction. Further thanks go to Johannes Schönberger as the main author of the open-source software COLMAP, which comprises both a structure-from-motion and a multi-view stereo pipeline for image-based 3D reconstruction.

In fetg grond engraziament va a mia famiglia sco era a mintgin che ha susteniu e motivau mei duront quei liung temps!