TECHNICAL UNIVERSITY OF MUNICH
TUM School of Engineering and Design
Photogrammetry and Remote Sensing

Geometric calibration of front cameras in vehicles for road scene acquisition

Alexander Hanel

Dissertation

2022

TECHNISCHE UNIVERSITÄT MÜNCHEN
TUM School of Engineering and Design

Geometric calibration of front cameras in vehicles for road scene acquisition

Alexander Hanel

Vollständiger Abdruck der von der TUM School of Engineering and Design der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr.phil.nat. Urs Hugentobler
Prüfer der Dissertation:	1. Prof. Dr.-Ing. Uwe Stilla
	2. Prof. Dr.-Ing. Boris Jutzi Karlsruher Institut für Technologie (KIT)

Die Dissertation wurde am 19.09.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 21.11.2022 angenommen.

Abstract

On-board forward-looking cameras are often used for environment perception for advanced driver assistance systems and for autonomous driving. In order to determine geometric quantities like the distance to a preceding car, calibration of these cameras is necessary. As forward-looking cameras are normally mounted inside the car, an influence of the windshield on the imaging geometry is imposed, which creates a special challenge for calibration. In addition, it can be assumed that changing camera properties over vehicle lifetime lead to errors in the determined geometric quantities. Therefore, methods for calibration of forward-looking cameras are proposed and investigated in this thesis.

To investigate the influence of the windshield, test field calibration of a stereo camera system is carried out. A flexible 3d test field is used and an approach for datum definition by free adjustment for mono cameras is extended for the stereo camera system. The camera orientation and their uncertainties are estimated by bundle adjustment.

To cope with changes in the camera properties during vehicle lifetime, a method for self-calibration with reference points obtained from the road scene is proposed and investigated. Remarkable points at the boundary of traffic signs serve as reference points. Their pixel and object coordinates are obtained by means of semantic segmentation, boundary detection and depth estimation. Thereby, scene knowledge is obtained by deep learning and the camera orientation is estimated by bundle adjustment with reference points from traffic signs of three different shapes.

As alternative to self-calibration with reference points from traffic signs, a method for self-calibration with reference points obtained from arbitrary objects in the road scene is proposed and investigated. Some of these points, like those on moving cars, are inappropriate for calibration. By means of semantic segmentation, undesired image regions are excluded and the matching of image points is improved. In addition, the vehicle trajectory obtained from GPS is evaluated by a special vehicle movement model for better metric scaling.

The proposed methods are evaluated with image sequences showing a test field and road scenes in a suburban and urban environment, respectively. The results of test field calibration show a statistically significant influence of the windshield on the camera orientation parameters. Self-calibration with traffic signs shows best results with triangular shaped traffic signs. The use of semantic segmentation improves self-calibration with points from arbitrary objects.

Kurzfassung

Zur Erfassung der Umgebung für Fahrerassistenzsysteme und zum autonomen Fahren werden in modernen Fahrzeugen häufig Frontkameras eingesetzt. Um geometrische Größen wie beispielsweise die Distanz zu einem vorausfahrenden Fahrzeug bestimmen zu können, müssen diese Kameras kalibriert werden. Da Frontkameras in der Regel im Fahrzeuginneren hinter der Windschutzscheibe montiert sind und die Windschutzscheibe die Abbildungsgeometrie beeinflusst, ergibt sich für die Kalibrierung eine besondere Herausforderung. Weiter ist davon auszugehen, dass sich über die Fahrzeuglebensdauer die Kameraeigenschaften verändern und zu Fehlern bei der Bestimmung geometrischer Größen führen. Daher werden in dieser Arbeit Methoden vorgeschlagen und untersucht, um die Kalibrierung von Frontkameras zu verbessern.

Zur Untersuchung des Einflusses der Windschutzscheibe wird eine Testfeldkalibrierung eines Stereokamerasystems durchgeführt. Dazu wird ein Ansatz mit einem beweglichen 3D-Testfeld verwendet und ein Verfahren zur Datumsdefinition per freier Ausgleichung für eine Monokamera für das Stereokamerasystem erweitert. Die Kameraorientierung und deren Genauigkeiten werden durch Bündelblockausgleichung geschätzt.

Um Veränderungen der Kameraeigenschaften während der Fahrzeuglebensdauer entgegenzuwirken, wird ein Verfahren zur Selbstkalibrierung mit Referenzpunkten aus der Straßenszene vorgeschlagen und untersucht. Als Referenzpunkte dienen markante Punkte am Rahmen von Verkehrszeichen, deren Pixel- und Objektkoordinaten mit Hilfe von semantischer Segmentierung, Kantendetektion und Tiefenschätzung bestimmt werden. Dabei wird Wissen über die Szene mittels Deep Learning gewonnen und Referenzpunkte von Verkehrszeichen dreier Formen zur Schätzung der Kameraorientierung mittels Bündelblockausgleichung verwendet.

Alternativ zur Selbstkalibrierung mit Referenzpunkten von Verkehrszeichen wird ein Verfahren vorgeschlagen, bei dem Referenzpunkte von beliebigen Punkten in der Straßenszene gewonnen werden. Dabei treten auch Punkte wie auf bewegten Fahrzeugen auf, die für eine Kalibrierung ungeeignet sind. Mit Hilfe von semantischer Segmentierung werden Bildbereiche ausgeschlossen und die Zuordnung von Bildpunkten verbessert. Zusätzlich wird die per GPS aufgenommene Fahrzeugtrajektorie für eine bessere metrische Skalierung durch ein spezielles Bewegungsmodell für Fahrzeuge ausgewertet.

Zur Evaluierung der Verfahren werden Bildsequenzen eines Testfeldes und natürlicher Szenen im vorstädtischen und städtischen Bereich verwendet. Die Ergebnisse der Testfeldkalibrierung zeigen, dass die Windschutzscheibe einen statistisch signifikanten Einfluss auf die Orientierungsparameter hat. Für die Selbstkalibrierung mittels Verkehrszeichen zeigt sich, dass sich besonders dreieckförmige Verkehrszeichen eignen. Bei der Selbstkalibrierung mit beliebigen Punkten verbessert die semantische Segmentierung das Kalibrierungsergebnis.

Contents

List of Abbreviations

Abbreviation	Description	Page
BRIEF	Binary robust independent elementary features	37
COLMAP	Framework for SfM	66
CV	Computer vision	21
DOF	Degree of freedom	33
EO	Exterior orientation	21
GPS	Global positioning system	3
HSV	Hue, saturation, value (color space)	54
IMU	Inertial measurement unit	35
IO	Interior orientation	21
LiDAR	Light detection and ranging	15
MAD	Median absolute deviation	72
MCS	Model coordinate system	45
MT	Mask type	68
MVS	Multi view stereo	31
ORB	Oriented FAST and rotated BRIEF	32
PCS	Pre-determined coordinate system	45
RANSAC	Random sample consensus	38
RGB	Red - Green - Blue	30
RGB-D	RGB and depth	31
RO	Relative orientation	21
SFE	Semantic feature extraction	68
SfM	Structure-from-motion	31
SIFT	Scale-invariant feature transform	32
SLAM	Simultaneous localization and mapping	31
SM	Semantic feature matching	68
SURF	Speeded up robust features	32
TF	Test field	45
USfM	Framework for uncertainty estimation for SfM	66
VO	Visual odometry	31
VMM	Vehicle motion model	68

List of Figures

List of Tables

1 Introduction

1.1 Automotive vision for assistance systems and autonomous driving

Observing the road scene environment around a vehicle is important for many advanced driver assistance systems and in particular on the way to autonomous driving. Different types of sensors for environment perception may be installed in modern vehicles [Winner et al., 2015; Ziebinski et al., 2016]. Ultrasonic, radar, LiDAR sensors or cameras are among the most common ones, covering complementary the distance range from few centimeters up to a few hundred meters and hence serve for applications like parking assistance, obstacle warning or adaptive cruise control [Zhang et al., 2014; Hanel et al., 2018]. Among the mentioned sensors, cameras are cheap and provide high-resolution data and hence are widely used in both mass-produced and research vehicles [Rosebrock & Wahl, 2012; Houben, 2014; Janai et al., 2017; Borgmann et al., 2018].

Cameras in cars either work in the visible spectrum for daylight applications [Zhang et al., 2014; Janai et al., 2017] or in the near-infrared and thermal infrared spectrum for night-time applications [Dong et al., 2007; Ge et al., 2009; Herrmann et al., 2018]. For environment observation, especially forward-looking visible-spectrum cameras recording the upcoming driveway of vehicles can be seen as most important, which are often installed in a mono camera or a stereo camera setup [Dang et al., 2009; Enzweiler & Gavrila, 2009; Keller et al., 2011]. For both setups, the image processing and computer vision tasks needed for the aforementioned applications are similar: For instance, specific road scene objects like road markings, pedestrians or traffic signs need to be detected and recognized [Scheller et al., 2007; Bertozzi et al., 2010] or geometric quantities like the road width, the location, size or velocity of detected environment objects relative to the ego-car need to be determined [Broggi et al., 2001; Bellino et al., 2005; Scheller et al., 2007; Alvarez et al., 2014; Bhardwaj et al., 2018]. It may also be necessary to perform multi-sensor fusion [Geiger et al., 2012; Heng et al., 2014; Guindel et al., 2017] or even to obtain an entire 3d environment reconstruction [Janai et al., 2017]. For reliable use in cars, these tasks need to be performed with high accuracy [Ribeiro et al., 2006; Dubey, 2016].

As special challenge for automotive vision, various constraints resulting from mass-production processes, vehicle design or price requirements may limit the selection of cameras and lenses [Broggi et al., 2001; Rosebrock & Wahl, 2012; Guindel et al., 2017; Muhovic & Pers, 2020]. As one example, narrow-angle lenses and large stereo baselines up to one meter need to be chosen for forward-looking cameras as they allow to detect and measure small road scene objects at distances up to a few hundred meters [Stein et al., 2010], but their neat integration into vehicle design with a geometrically-stable mounting is complicated [Mentzer et al., 2017; Muhovic & Pers, 2020]. As another example, wide-angle or fisheye lenses can provide large fields of views for surround-view cameras, but at the cost of large distortions [Rosebrock & Wahl, 2012; Häne et al., 2017]. For research vehicles, other and additional challenges may arise, like the need to use detachable cameras that can be placed at different positions and angles at the car from time to time [Paula et al., 2014].

1.2 Automotive camera calibration for reliable automotive vision

For safe operation of advanced driver assistance systems and especially to resolve the challenges for automotive vision, automotive camera calibration is a key aspect [Broggi et al., 2001; Kluger et al., 2017]. Valid camera calibration parameters, i.e. typically the interior, exterior and, in the case of multi camera systems, the relative orientation, allow to localize the ego-car in its environment with high reliability, allow to get a consistent representation of the 3d environment around the car and allow to determine the mentioned geometric quantities even at a far range [Marita et al., 2006; Hansen et al., 2012; Knorr, 2018; Muhovic & Pers, 2020].

Same as automotive vision, so is also automotive camera calibration in particular faced with challenges. While for automotive forward-looking infrared cameras locations at the front bumper are favorable due to their spectral working range [Bertozzi et al., 2010], visible-range cameras are typically mounted inside the vehicle behind the windshield to protect them against environmental influences [Broggi et al., 2001; Gehrig, 2005; Dang et al., 2009; Franke et al., 2013; Livyatan & Berberian, 2017]. Hence as first challenge, in the presence of a glass windshield in the optical path between the camera and objects in front of the car, the important collinearity assumption is not fulfilled anymore due to refraction of the image rays at the air - glass transition [Maas, 2015b]. In particular at short distances in front of the vehicle and for vertically large objects, ignoring the effect of windshield refraction may heavily influence distance measurements, for instance [Lasaruk & Neralla, 2018; Verbiest et al., 2020]. This influence on the imaging geometry is reported to be "surprisingly large" [Lasaruk & Neralla, 2018; Verbiest et al., 2020]. Automotive stereo camera systems are in particular affected, as even a small error of just a few seconds in the relative orientation can remarkably alter the epipolar geometry and so decrease the quality of 3d environment reconstruction, distance or velocity estimation, especially for objects far away from the ego-car [Marita et al., 2006; Winner et al., 2015; Ling & Shen, 2016]. Additionally, the often large stereo baselines in automotive stereo camera systems increase the effect of such errors [Häne et al., 2017; Mentzer et al., 2017; Zabatani et al., 2017; Muhovic & Pers, 2020]. For the reasons mentioned, it seems obvious that the windshield refraction should be considered especially for high-quality test field calibration of automotive stereo camera systems [Geiger et al., 2012; Fraser, 2013].

Other automotive-specific problems may arise from the conditions on the road. Hence as second challenge, cameras must be operable over the entire vehicle lifetime, which requires the estimated camera orientation parameters to stay reliable [Bertozzi et al., 2010]. Mechanical, thermal and aging effects in the car could cause decalibration, i.e. changes in the geometry of the camera that lead to a drift in the orientation parameters, thus making previous camera calibration invalid and requiring re-calibration [Pflug et al., 2013; Gopaul et al., 2016; Mentzer et al., 2017; Rehder et al., 2017]. It is reported that even moderate effects may cause remarkable changes in the geometry [Broggi et al., 2001; Muhovic & Pers, 2020]. Vibrations in the vehicle from driving at higher speed, bad road conditions or collisions are just a few examples for the many anticipatable or not anticipatable sources for mechanical effects [Gopaul et al., 2016; Mentzer et al., 2017, 2019]. Ambient temperature variations or heating-up of cameras are just two sources for thermal effects [Gopaul et al., 2016; Adamczyk et al., 2018]. Also environmental conditions on the road that are different from production or research facilities where initial calibration has been done may cause decalibration [Cannelle et al., 2012] and so impose an additional challenge that needs to be resolved.

Therefore, solely calibrating automotive cameras once, often at the end of the production line with test fields, is not considered as sufficient for reliable use over vehicle lifetime [Ruland et al., 2010; Winner et al., 2015]. With even more emphasis, Bodis-Szomoru et al. [2008] and Heng et al. [2014] see repeated validity checks of previously estimated camera orientation parameters

or repeated re-calibration as indispensable. As solution, self-calibration of cameras carried out while driving on the road can ensure the needed quality of calibration parameters to reliably operate automotive vision applications over lifetime [Scheller et al., 2007; Musleh et al., 2014], even without the need for any labor- and cost-intensive efforts for test field calibration [Mueller & Wuensche, 2017; Knorr, 2018]. Therefore, reference points for calibration need to be obtained from the road scene, wherefore often road markings shown in the lower image half are utilized. As a result, the estimated parameters are only valid for this image half, motivating the use of complementary road scene objects typically shown in the upper image half, like traffic signs. While a certain type of objects typically occurs only in a small number and at certain locations of the scene, may a high number of reference points at various locations be obtained alternatively from arbitrary road scene objects. Though, the movement of objects like cars or reflective surfaces like building windows may render reference points on these objects invalid. If such points are used for calibration, a decrease in the quality of the estimated orientation parameters has to be expected, wherefore it is desirable to exclude them.

1.3 Research questions

In this thesis, the following research questions on the two mentioned automotive-specific challenges are addressed.

- How does a vehicle windshield in the optical path between a forward-looking on-board stereo camera system and a calibration test field influence the parameter values, standard deviations and correlations of the interior, relative and exterior orientation parameters estimated by test field calibration based on bundle adjustment in a set of experimental cases covering two kinds of test fields, different camera models and different parametrizations of stereo constraints?

- Which types of traffic signs are most appropriate to derive reference points from by deep learning-based computer vision for self-calibration with a sequence of road scene images taken with a forward-looking on-board mono camera?

- How can semantic road scene knowledge and vehicle motion models be integrated into a structure-from-motion pipeline to improve self-calibration of a forward-looking on-board mono camera with a series of road scene images?

1.4 Contributions

The key contributions of this thesis are as follows.

First, a method for stereo camera calibration with test fields is proposed that allows to jointly use reference points from two non-rigid 2d test fields as well as reference points from coded and uncoded reference marks by establishing point associations based on multiple similarity transformations. An existing approach for datum definition by free adjustment for mono cameras is extended for stereo camera systems. Investigations on the influence of the vehicle windshield on camera calibration are carried out with the proposed method. While in previous work the presence of an influence is shown, special emphasis of the investigation is put on the uncertainties and correlations of the estimated camera orientation parameters. Furthermore, the investigation comprises comparative evaluation of experiments with two kinds of test fields, different stereo constraints, camera models and bundle adjustment properties.

Second, a method for camera self-calibration is proposed whereby reference information is obtained from certain road scene objects. While in previous work often road markings are used, the proposed method relies on remarkable points at the boundary of traffic signs. The pixel and object coordinates of these points are calculated by exploiting scene knowledge. By utilizing deep learning-based semantic segmentation, boundary detection and single image depth estimation to obtain scene knowledge, the need for additional data sources like GPS or IMU for self-calibration can be avoided. Furthermore, the effect of three different shapes of traffic signs, two methods for semantic segmentation and two methods for depth estimation on the calibration results is investigated with test data from urban and suburban road scenes.

Third, a method for camera self-calibration with reference points obtained from arbitrary objects in the road scene is proposed. While in previous work only generic outlier removal has been applied, scene knowledge is obtained from semantic segmentation used in the proposed method to exclude points from automotive-specific undesired objects like moving cars or reflecting windows. While other previous work either addresses only feature extraction or feature matching, the proposed method (i) applies masks to exclude image parts with undesired semantic classes during extraction of SIFT features and (ii) restricts feature matching to points belonging to the same semantic class. Furthermore, while previous work considers only moving objects as undesired, uses only synthetic images for testing or relies on visual SLAM, the proposed method also considers semantic object classes with inappropriate surfaces (e.g. poor textures like tarmac, reflecting surfaces like building windows) as undesired. Furthermore, it relies on a typically better performing structure-from-motion approach for self-calibration. Additionally, a vehicle trajectory obtained from GPS is refined by Kalman filtering with a special vehicle movement model for a better metric scaling in the structure-from-motion approach. The method is evaluated with real image sequences from suburban and urban scenes, whereby the effect of masks created from different combinations of semantic classes, the benefits of restricting feature matching and filtering with the vehicle motion model are investigated.

Parts of this thesis have been published in the following papers:

- [Hanel et al., 2016] Hanel A, Hoegner L, Stilla U (2016) Towards the influence of a car windshield on depth calculation with a stereo camera system. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLI-B5: 461-468.
- [Hanel & Stilla, 2017] Hanel A, Stilla U (2017) Structure-from-motion for calibration of a vehicle camera system with non-overlapping fields-of-view in an urban environment. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-1/W1: 181-188.
- [Hanel & Stilla, 2018] Hanel A, Stilla U (2018) Iterative calibration of a vehicle camera using traffic signs detected by a convolutional neural network. In: International Conference on Vehicle Technology and Intelligent Transport Systems: 187-195.
- [Hanel et al., 2018] Hanel A, Kreuzpaintner D, Stilla U (2018) Evaluation of a traffic sign detector by synthetic image data for advanced driver assistance systems. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2: 425-432.
- [Hanel & Stilla, 2019a] Hanel A, Stilla U (2019a) Evaluation of iterative calibration of vehicle cameras using reference information from traffic signs. In: Donnellan B, Klein C, Helfert M, Gusikhin O (eds) Smart Cities, Green Technologies and Intelligent Transport Systems.: Springer, CCIS, 992, 244-265.

- [Hanel & Stilla, 2019b] Hanel A, Stilla U (2019b) Semantic road scene knowledge for robust self-calibration of environment-observing vehicle cameras. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W16: 103-110.
- [Hanel et al., 2019] Hanel A, Sudi P, Pfenninger S, Steinbach E, Stilla U (2019) Filter-based pose estimation for electric vehicles relative to a ground-based charging platform using on-board camera images. In: Kersten TP (ed) Wissenschaftlich-Technische Jahrestagung der DGPF, 28, 54-67.

1.5 Structure of the thesis

An introduction into the relevant basics and definitions of camera calibration, adjustment theory and computer vision are given in Chapter 2. The state of the art in automotive camera calibration is discussed in Chapter 3. In Chapter 4, the method for test field calibration of a stereo camera system for investigation of the windshield effect is described, while Chapter 5 addresses the method for camera self-calibration with reference information derived from traffic signs. Chapter 6 comprises the method for camera calibration by semantic structure-from-motion. Descriptions of the test datasets and experiments are given in Chapter 7. Results and discussion are covered in Chapter 8. The thesis concludes with answers on the research questions and an outlook given in Chapter 9.

2 Basics and definitions

This chapter covers relevant basics of camera calibration, adjustment theory and computer vision, beginning with geometric camera calibration (Subsection 2.1.1). In the following, the differences of camera calibration in photogrammetry and computer vision (Subsection 2.1.2) as well as the differences between test field and self-calibration (Subsection 2.1.3) are addressed. Afterwards, camera and distortion models are introduced (Subsection 2.1.4). Important aspects of automotive camera calibration (Subsection 2.1.5) and calibration algorithms (Subsection 2.1.6) are covered. Then, basic aspects of adjustment theory are addressed (Subsections 2.2.1 and 2.2.2), followed by details on the important collinearity equations and on the topic of datum definition (Subsections 2.2.3 and 2.2.4). In the last part of this Chapter, an introduction into the computer vision topics of object detection and segmentation (Subsection 2.3.1) as well as of 3d reconstruction and localization is given (Subsection 2.3.2).

2.1 Camera calibration

2.1.1 Geometric camera calibration

Geometric camera calibration is required to obtain accurate metric information from images [Pollefeys & Van Gool, 1997; Remondino & Fraser, 2006; Luhmann et al., 2016] and aims at determining the interior (IO), relative (RO) or exterior orientation (EO) parameters; the approaches proposed in this thesis address either one or more of these three parameter groups. Various definitions for the terms *relative orientation* and *exterior* or *extrinsic* orientation can be found in the literature: They can refer to the orientation between two cameras [Stein et al., 2010; Mentzer et al., 2017], the orientation between cameras and other types of sensors [Domhof et al., 2019] or the orientation between a camera and the vehicle [Broggi et al., 2001; Catala-Prat et al., 2006]. In this thesis, *relative orientation* refers to the orientation between two cameras and *exterior orientation* refers to the orientation of a camera in a higher-level coordinate system, like an *object* or *vehicle* coordinate system. The term *extrinsic orientation* is not used. With focus on the type of reference information, camera calibration can be categorized into (i) laboratory calibration, (ii) test field calibration and (iii) self-calibration [Luhmann et al., 2006; Kraus, 2007; Förstner & Wrobel, 2016]. Laboratory calibration relies on using optic measures, e.g. a collimator, for calibration. Due to the required special laboratory equipment and the high effort, this approach is feasible only if the required special conditions can be met and very high accuracy is demanded. Test field calibration relies on test fields with known reference information, like points with known 3d object coordinates [Sturm & Maybank, 1999] or plumb lines [Brown, 1971], and typically a set of images showing this reference information to determine the desired camera orientation parameters. Self-calibration relies either on test fields without a priori known reference information or on a sufficient number of reference information that can be identified from the scene [Luhmann et al., 2016]. Hereby, the object coordinates of the reference points are obtained during calibration. Self-calibration is considered to be the most general and simple approach [Luhmann et al., 2006; Förstner & Wrobel, 2016], and as an "integral and routinely

applied operation" within photogrammetry, especially "in close-range measurement" [Remondino & Fraser, 2006]. As confirmed by Fraser [2013], self-calibration with bundle adjustment can be seen as the current norm in close-range photogrammetry. Nevertheless, stand-alone calibration with test fields "has again emerged as an important issue in close-range photogrammetry", for example in cases where the geometry of the image network does not provide enough support for a robust estimation of the camera orientation parameters by self-calibration [Remondino & Fraser, 2006]. Automotive calibration covered in this thesis is considered to belong to the field of close-range photogrammetry due to the type of used cameras and the terrestrial imaging configuration, in contrast to e.g. aerial photogrammetry. Both automotive test field and self-calibration are addressed. In addition to the descriptions above, other kinds of categorization are known as well [e.g Shih et al., 1996; Luhmann et al., 2006]; though, their discussion is out of the scope of this thesis. Equally, radiometric camera calibration [e.g. Mo et al., 2017] addressing the relation between image intensities and scene radiance [Li et al., 2017] is not covered in this thesis.

2.1.2 Camera calibration in photogrammetry and computer vision

Both the photogrammetry [e.g. Fraser, 1997; Remondino & Fraser, 2006; Luhmann et al., 2013] and the computer vision (CV) community [e.g. Tsai, 1987; Maybank & Faugeras, 1992; Zhang, 2000] address camera calibration, but common differences in the approaches and objectives can be identified. According to Fraser [1997], calibration in the CV community often focus on minimal geometric information, i.e. the least-possible number of images or the least-possible number of reference points, which may lead to scene-dependent solutions and highly correlated camera orientation parameters [Fraser, 2013]. Simultaneous calibration during measurement campaigns seems to be more common than a priori calibration to avoid decalibration caused by mechanical or thermal effects [Fraser, 1997; Remondino & Fraser, 2006]. Furthermore, calibration approaches from the CV community are often designed to be easy-to-use and fully automated, with error analysis not being in the focus [Shih et al., 1996]. In contrast, calibration in the photogrammetry community often concentrates on high quality [Luhmann et al., 2016] and on thorough result analysis [Fraser, 1997]. High quality might be visible by high parameter accuracy, strong image networks, high redundancy, a good initial guess and the use of complex Gauß-Markov optimization [Börlin & Grussenmeyer, 2014; Luhmann et al., 2016]. Even though simultaneous calibration approaches are used, photogrammetric calibration is often implemented as a priori calibration [Börlin & Grussenmeyer, 2014]. As it will be reflected by the following chapters, this thesis relies on aspects from both communities.

2.1.3 Test field calibration and self-calibration

Test field calibration relies on test fields providing a set of reference information, which is mostly points represented by reference marks. The reference information has to be known prior to calibration, hence the object coordinates of such points have to be determined by high-quality photogrammetry or tacheometry, for instance. The image coordinates are obtained during calibration by image processing tailored to the appearance of the reference marks. In close-range photogrammetry and computer vision, checkerboard patterns are a popular type of 2d test fields, where the checkerboard corners serve as reference points [Luhmann et al., 2016]. Alternatively, patterns of circular reference marks attached to bars or planes are popular as well [Vo et al., 2011; Schneider et al., 2017]. Multiple images taken in a suitable imaging geometry are required for calibration with 2d test fields, while at minimum a single image is sufficient for calibration with 3d test fields [Urban et al., 2015]. According to Luhmann et al. [2016], calibration with planar sets of reference points can easily lead to undesired high correlations between interior and exterior orientation parameters, wherefore 3d test fields are taken in the scope of this thesis.

In contrast, self-calibration employs various single-image and multi-image constraints that have to be extracted from the acquired scene [Ling & Shen, 2016]. Single-image constraints may originate from homographies between object space and the image (e.g. planar checkerboard test fields without known object coordinates) or from multiple vanishing points corresponding to orthogonal directions (e.g. along the edges of a rectangular building) [Liebowitz, 2001], for example. Multi-image constraints may originate from epipolar geometry (two images) [Maybank & Faugeras, 1992] and can be modeled for planar point sets by homographies [Miksch et al., 2010], or for non-planar sets by the essential matrix for cameras with known interior orientation [Bjorkman & Eklundh, 2002] or the fundamental matrix for cameras without known interior orientation [Faugeras et al., 1992]. Multi-image constraints may also rely on the trifocal tensor (three images) [Armstrong et al., 1996; Hartley, 1997a] or bundle adjustment (typically more than three images) [Fraser, 2013], for example. Other constraints may originate from projective geometry and allow to determine the interior orientation parameters based on recovery of geometric items whose projections stay fixed throughout an image sequence [Pollefeys & Van Gool, 1997]; examples are the image of the absolute conic [Faugeras et al., 1992] or the image of absolute dual quadric [Triggs, 1997]. While some methods assume all interior orientation parameters to be constant for successful recovery [e.g. Faugeras et al., 1992], others allow that a subset of these parameters may have varying values [Heyden & Astrom, 1997]. While some methods rely on certain motion patterns, for example pure translational [e.g. Dron, 1993], pure rotational [e.g. Hartley, 1997b] or pure planar motion [e.g. Armstrong et al., 1996], others are prone to critical motion patterns that could cause calibration to fail and hence need to be avoided [Hartley & Zisserman, 2003]. Within this thesis, bundle adjustment is used for the proposed methods, as plenty of reference information is available and there is no focus on calibration with a minimal geometric configuration. For all methods, the interior and relative orientation are assumed to be constant during acquisition of the calibration images.

2.1.4 Camera and distortion models

To model the imaging geometry, the theoretical assumption of an ideal central projection is made. In reality, it is violated by perturbations caused by lens distortions, chromatic aberration or non-planarity of the sensor surface, for example [Fraser, 1997; Förstner & Wrobel, 2016; Granshaw, 2020]. As insufficient modeling of the projection and perturbations are a typical source for calibration errors [Heikkila, 2000], an appropriate camera and distortion model is important for successful camera calibration. Various types of camera models are known: Probably most common is the pinhole camera model [Brown, 1971; Heikkila, 2000] which can serve for calibration of cameras with narrow-angle or even wide-angle lenses, if accompanied by appropriate lens distortion models [Kannala & Brandt, 2006]. Models for special cameras or lenses, like fisheye lenses [Kannala & Brandt, 2006] or omnidirectional cameras [Scaramuzza et al., 2006] exist. As such special cameras or lenses are not used by the proposed methods and experiments, further details on them are not covered here. Common lens distortion models allow for correction of radial-symmetric and tangential (decentering) distortion [Brown, 1971] as well as for affinity or shear [El-Hakim, 1986]. According to Bergamasco et al. [2013], so-called specific camera models as described up to here, provide easy-to-use, well-adapted models for certain cameras; however, they bear a trade-off between considering all perturbating effects and a low number of parameters that can be determined reliably during calibration. In contrast, so-called generic high-parameter camera models [e.g. Kannala & Brandt, 2006] allow to easily consider various projections and distortions and hence are suitable for several types of cameras and lenses at the same time [Guo-Qing Wei & Song De Ma, 1994; Rosebrock & Wahl, 2012]. As such generality is not needed for this thesis, a standard pinhole camera model with different lens distortion models is taken, modeling

either no distortions, modeling the typically stronger radial distortions or modeling both radial and tangential distortions (Table 2.1).

Table 2.1: Camera and distortion models determine the vector of interior orientation parameters $\boldsymbol{X}_I$ used for camera calibration. f represents the focal length, c the principal point coordinates, k the radial distortion and p the tangential distortion parameters, each in x- and y-direction of the image coordinate system.

Name	*NONE*	*RADIAL*	*BOTH*
$\boldsymbol{X}_I$	$(f_x \quad f_y \quad c_x \quad c_y)$	$(f_x \quad f_y \quad c_x \quad c_y \quad k_1 \quad k_2)$	$(f_x \quad f_y \quad c_x \quad c_y \quad k_1 \quad k_2 \quad p_1 \quad p_2)$

In the vector $\boldsymbol{X}_I$ of interior orientation parameters, f_x and f_y represent the focal length for the x- and y-axis of the image coordinate system. Note that the focal length differs from the principal distance depending on the camera focus [Granshaw, 2020]; as it is more common, the term focal length is used in the further course. c_x and c_y represent the principal point coordinates in x- and y-direction. k_1 and k_2 represent radial distortion parameters compensating for radial-symmetric effects, and p_1 and p_2 represent tangential distortion parameters compensating for effects of lens decentering. These models have been selected based on their common use [e.g. Förstner & Wrobel, 2016; Luhmann et al., 2016; Schönberger & Frahm, 2016; OpenCV, 2017; Polic et al., 2018]. Note that the described parametrization of the distortion models is taken from the computer vision community, as to the author's knowledge more work for automotive camera calibration originates from this community. In the photogrammetry community, radial-symmetric distortion parameters are often alternatively referred to as A_1 etc., and tangential distortion parameters as B_1 etc. [Luhmann et al., 2016]. More important to acknowledge, there are different opinions in the communities on the mathematical formulation of distortion models that could result in small differences compared to the formulations used in this thesis, for example with regard to the use of an additional radius of zero-crossing or with regard to the series expansion for radial-symmetric distortion [Luhmann et al., 2006; Förstner & Wrobel, 2016]. Other perturbations than lens distortions are not considered in this thesis. According to Luhmann et al. [2016], such perturbations may be worth to consider in extended camera models designed for special cameras or conditions.

2.1.5 Automotive camera calibration

Automotive camera calibration is used to estimate either the interior [Houben, 2014; Keivan & Sibley, 2015; Hanel & Stilla, 2018], the exterior [Ruland et al., 2010; Heng et al., 2014] or in the case of multi-camera systems the relative orientation. Other approaches aim at simultaneously estimating two or more types of orientation [e.g. Heng et al., 2013]. Thereby, calibration provides the mapping between images of automotive cameras and the road scene environment [Houben, 2014; Häne et al., 2017] and with a calibrated automotive camera, angle, distance or velocity measurements between the ego-car and environment objects become possible. In this thesis, the interior orientation is estimated by all proposed methods, and additionally the relative and exterior orientation are estimated by one method. In the automotive domain as well as in the proposed methods, calibration of on-board mono cameras [e.g. Miksch et al., 2010] or multi-camera systems [e.g. Broggi et al., 2001] are addressed. As besides normal-angle lenses also wide-angle and fisheye lenses are common for automotive cameras [Rosebrock & Wahl, 2012; Heng et al., 2013], the remarkable lens distortions may not be neglected during calibration. Most work on automotive camera calibration addresses forward-looking camera systems [e.g. Dang et al., 2009; Hanel et al., 2016], while other addresses downward-looking camera systems [e.g. Pliefke, 2013]. Camera calibration in the automotive domain addresses also multi-sensor systems, like a combination of camera and LiDAR or radar [Schöller et al., 2019; Geiger et al., 2012; Levinson & Thrun, 2013; Schöller et al., 2019], or "off-board" cameras, e.g. stationary road surveillance

cameras [Ismail et al., 2010; Brown et al., 2015]. The later two kinds of cameras and systems are not covered in this thesis.

Automotive camera calibration is done as either test field calibration [Broggi et al., 2001; Geiger et al., 2012; Hanel et al., 2016; Cordts et al., 2016] or self-calibration [Stein et al., 2010; Mueller & Wuensche, 2017; Rehder et al., 2017]. Both so-called "offline" and "online" calibration are employed [Cannelle et al., 2012; Ishikawa et al., 2018] and used by the methods in this thesis. Hereby, online calibration takes place in the area of application and typically relies on a high number of reference points obtained while driving on the road [Cannelle et al., 2012; Dlugosz et al., 2019]. In contrast, offline calibration does not take place in the area of application, i.e. often in a factory [Dlugosz et al., 2019] or in research facilities [Geiger et al., 2012]. Test field calibration is often used as initial calibration [e.g. Gil et al., 2018a; Lasaruk & Hachfeld, 2019], for example at the end of camera production or at the end of the car production line is mass-production facilities [Ruland et al., 2010; Winner et al., 2015; Lasaruk & Hachfeld, 2019]. Such initial calibration enables or facilitates subsequent re-calibration [Houben, 2014]. As mechanical solutions to avoid decalibration over time are difficult to realize [Lasaruk & Hachfeld, 2019], can single end-of-line calibration or laboratory calibration not be seen as sufficient to provide and ensure calibration parameters that are valid over vehicle lifetime. Repeated re-calibration, or *continuous calibration* [Dang et al., 2009], can serve to overcome temporal decalibration by a validity check of previously estimated calibration parameter values [Marita et al., 2006; Szczepanski, 2019] or by an update to these values, if necessary [Broggi et al., 2001]. In particular, this is relevant for stereo cameras [Winner et al., 2015]. Typically, repeated re-calibration is performed as online calibration, i.e. the calibration parameters are estimated with reference information derived from the ego-car or the road scene environment while driving the car on the road [Mueller & Wuensche, 2017; Knorr, 2018; Mentzer et al., 2019; Paone et al., 2019]. Mostly, online calibration is accomplished by self-calibration [Bellino et al., 2005; Heng et al., 2013; Rehder et al., 2017; Zheng & Zhao, 2017], which does also apply to two of the proposed methods.

Parallels between automotive camera calibration and camera calibration in close-range photogrammetry or computer vision exist obviously with regard to the estimated parameters, i.e. the interior, relative or exterior orientation parameters [Kruger et al., 2004; Scheller et al., 2007; Winner et al., 2015; Gopaul et al., 2016]. Parallels exist also with regard to the use of similar algorithms [Broggi et al., 2001; Heng et al., 2013; Lasaruk & Neralla, 2018]. Differences can be mainly found in the use of automotive-specific test fields or road scene-specific reference information, as it will be further discussed in Chapter 3. As specific property of automotive camera calibration, special precautions have to be taken in particular for mass-produced cars due to the limited computational power that is available. For example, global bundle adjustment processing the entire available image sequence may be replaced by local bundle adjustment processing only subsets of the image sequence [Rehder et al., 2017], the optimization may be performed in a reduced-order setup [Dang et al., 2009] or computationally light-weight recursive approaches like Kalman filtering may be utilized [Mueller & Wuensche, 2017]. But as the methods proposed in this thesis are intended for and tested with a research setup, limitations in computational power are not considered in the design of the calibration algorithms.

2.1.6 Calibration algorithms

Calibration can be realized by various types of algorithms. Classic approaches often incorporate a two-step algorithm with an initial linear closed-form solution followed by refinement based on non-linear optimization [e.g. Heikkila & Silven, 1997; Häne et al., 2017]. For example, a linear solution can be obtained by direct linear transformation (DLT) [Abdel-Aziz & Karara, 1971]. Often linear solutions base on simplified models, e.g. without considering distortions.

Optimization often utilizes in the first step geometrically meaningless and linear algebraic distance measures derived from constraints that can be easily minimized [Zhang & Pless, 2004]. Geometric distance measures, e.g. the re-projection error, are often used in the subsequent refinement step to improve previously obtained results to achieve higher accuracy [Zhang & Pless, 2004; Rodehorst et al., 2008; Dang et al., 2009]. Bundle adjustment as one of the most comprehensive optimization approaches often forms the last step of a calibration algorithm [Liebowitz, 2001; Hartley & Zisserman, 2003]. With bundle adjustment, high accuracy and low re-projection errors down to a fraction of one pixel can be achieved [Rosebrock & Wahl, 2012], but at the cost of larger processing times [Ling & Shen, 2016] compared to other algorithms or constraints and at the need of a known initial guess [Okouneva, 2017]. According to Dang et al. [2009], the accuracy achievable with other algorithms or constraints is not comparable to bundle adjustment. Recursive algorithms, like the Kalman filter [Hansen et al., 2012], allow iteratively integrating new measurements over time into the optimization process and therefore are particularly suitable for continuous calibration [Dang et al., 2009]. Obviously, recently calibration approaches based on deep learning have been published [Bogdan et al., 2018; Gil et al., 2019]. With deep learning, beneficial properties similar to classic calibration approaches can be achieved, like single shot calibration or no need for test fields [Bogdan et al., 2018; Hold-Geoffroy et al., 2018]. But as deep learning for camera calibration is a new field of research, the approaches currently bear certain problematic properties as well: As important aspect, generalization and robustness of the approaches with regard to scenes that are different from the training data can be questionable [Gil et al., 2019]. Bogdan et al. [2018] and Hold-Geoffroy et al. [2018] state that currently only low-resolution images are supported, creating a training dataset with ground truth interior orientation parameter values is necessary and that images with motion blur, overexposure or images taken with rolling shutter cameras often show unreliable results. The same authors add that also images taken in nadir direction may be problematic, in addition to the lower accuracy compared to classic calibration approaches. Interestingly, Hold-Geoffroy et al. [2018] have revealed that their model seems to learn "semantically meaningful vanishing lines, making parallels with geometrically-based auto-calibration techniques". But as no advantage of deep learning is seen, the methods proposed in this thesis rely on classic camera calibration with bundle adjustment, same as other recent methods for automotive camera calibration [e.g. Okouneva, 2017; Lasaruk & Hachfeld, 2019].

2.2 Adjustment theory

Bundle adjustment is an optimization method to simultaneously estimate the interior, potentially relative, and exterior orientation parameters as well as the 3d object coordinates of the reference points as unknown parameters in a statistically optimal manner [Förstner & Wrobel, 2016]. Typical observations are pixel or image coordinates of the reference points for calibration that are shown in multiple images and, depending on the calibration setup and mathematical model, additionally the object coordinates of the reference points. The functional basis of bundle adjustment is typically defined by the collinearity equations. Especially in photogrammetry, the optimization is realized by non-linear least squares adjustment with the Gauß-Markov or Gauß-Helmert model, wherefore the basics will be described in this section.

2.2.1 Adjustment basics

Each mathematical model that is used to solve parameter estimation problems by non-linear least squares consists of a functional and a stochastic model [Förstner & Wrobel, 2016]. The functional

model of the first approach, called Gauß-Markov model, is defined in its basic unconstrained case as

$$\boldsymbol{b}_1 + \hat{\boldsymbol{v}}_{\boldsymbol{b}_1} = \boldsymbol{f}_1(\hat{\boldsymbol{x}}) \tag{2.1}$$

with $\boldsymbol{b}_1$ being observations, $\hat{\boldsymbol{v}}_{\boldsymbol{b}_1}$ being their residuals, $\boldsymbol{x}$ being the unknown parameters, the circumflex symbol ("hat") denoting estimated quantities and $\boldsymbol{f}_1(...)$ being one or more functions relating the observations with the parameters. The functional model of the second approach, called Gauß-Helmert model, is defined in its basic unconstrained case as

$$\boldsymbol{f}_1(\boldsymbol{b}_1 + \hat{\boldsymbol{v}}_{\boldsymbol{b}_1}, \hat{\boldsymbol{x}}) = \boldsymbol{0} \tag{2.2}$$

In the case of hard (crisp) constraints, additional dependencies between the parameters can be modeled that are enforced strictly during optimization. Therefore, the respective functional model is extended by

$$\boldsymbol{f}_2(\hat{\boldsymbol{x}}) = \boldsymbol{0} \tag{2.3}$$

with $\boldsymbol{f}_2(...)$ being one or more constraining functions. In contrast, in the case of weak (soft) constraints, where it is allowed that after optimization residuals greater than zero remain for these constraints, the functional model is extended by fictional observations of type $\boldsymbol{b_1}$ and by additional constraining functions of type $\boldsymbol{f}_1(...)$. The influence of these constraints is controlled by observation weights defined in the same way as observation weights for real observations. Fictional observations may be the expected relative position or rotation between two cameras, for example, and the corresponding constraining functions may calculate this position or rotation from the exterior orientation parameters of both cameras that are modeled as unknown parameters. As stochastic model for both the Gauß-Markov and Gauß-Helmert model, normally-distributed observations with covariance matrix $\boldsymbol{K}_{bb} = \sigma_0^2 \cdot \boldsymbol{Q}_{bb}$ are assumed, with σ_0^2 being the variance factor a priori and $\boldsymbol{Q}_{bb}$ being the weight coefficient matrix. The earlier controls the overall weight level for all observations, while the later contains the variances of the observations on the main diagonal, and the covariances between observations on the secondary diagonals and so allows to determine weight ratios between different observations. Due to the common non-linearity of the functional model, an initial guess has to be selected for the unknown parameters to start the iterative estimation process with. In this iterative process, the unknown parameter values are updated by minimizing a cost function based on the weighted linearized functional model. Typically, first order Taylor series expansion is utilized for linearization. The process is stopped when a convergence criterion has been met, for example if the updates of the unknown parameter values fall below a given threshold. According to Förstner & Wrobel [2016], the prevalent estimation problem guides the decision for one of the two models.

2.2.2 Uncertainty of observations and unknown parameters

The uncertainty of observations and estimated unknown parameter values plays an important role to evaluate camera calibration. For optimization with the Gauß-Markov and Gauß-Helmert model, uncertainties can be calculated after convergence and are typically described by the variance factor a posteriori, the covariance matrix of unknown parameters and the covariance matrix of observations. The estimated standard deviations of the unknown parameters, which are an important aspect of evaluation in this thesis (Chapter 8), can be extracted from the corresponding covariance matrix. Other measures, like correlation coefficients, can be obtained by calculation using elements from these matrices. Uncertainty measures could be also obtained by other approaches, for example by error propagation [e.g. Hartley & Zisserman, 2003] or by Kalman filtering [Kalman, 1960], but this is not done for this thesis.

2.2.3 Collinearity equations

The collinearity equations are basic photogrammetric equations relating 2d pixel or image coordinates of observed points, like reference points for camera calibration, with the interior and exterior camera orientation as well as the 3d object coordinates of these points [Kraus, 2007]. The non-linear collinearity equations are often used as functional model for bundle adjustment and are defined as

$$f_{x,i,j,k} : x_{i,j,k} = c_{x,j} - f_{x,j} \cdot \frac{R_{1,1,j,k} \cdot (X_i - X_{0,j,k}) + R_{2,1,j,k} \cdot (Y_i - Y_{0,j,k}) + R_{3,1,j,k} \cdot (Z_i - Z_{0,j,k})}{R_{1,3,j,k} \cdot (X_i - X_{0,j,k}) + R_{2,3,j,k} \cdot (Y_i - Y_{0,j,k}) + R_{3,3,j,k} \cdot (Z_i - Z_{0,j,k})} \quad (2.4)$$

$$f_{y,i,j,k} : y_{i,j,k} = c_{y,j} - f_{y,j} \cdot \frac{R_{1,2,j,k} \cdot (X_i - X_{0,j,k}) + R_{2,2,j,k} \cdot (Y_i - Y_{0,j,k}) + R_{3,2,j,k} \cdot (Z_i - Z_{0,j,k})}{R_{1,3,j,k} \cdot (X_i - X_{0,j,k}) + R_{2,3,j,k} \cdot (Y_i - Y_{0,j,k}) + R_{3,3,j,k} \cdot (Z_i - Z_{0,j,k})} \quad (2.5)$$

given for reference point i, for camera j and for image k. The collinearity equations can be extended by the correction terms Δx and Δy, often used to model image distortions so that

$$x'_{i,j,k} = x_{i,j,k} + \Delta x_{i,j,k} \quad (2.6)$$

$$y'_{i,j,k} = y_{i,j,k} + \Delta y_{i,j,k} \quad (2.7)$$

with $x_{i,j,k}$, $y_{i,j,k}$ describing the undistorted, but unobservable points and $x'_{i,j,k}$, $y'_{i,j,k}$ describing the distorted and observable points. Additionally, the object coordinates of the reference points can be also modeled as observations to consider them with a realistic observation weight. Then, for each reference point there will be three additional functional equations defined as

$$X_i + \hat{v}_{X_i} = \hat{X}_i \quad (2.8)$$

$$Y_i + \hat{v}_{Y_i} = \hat{Y}_i \quad (2.9)$$

$$Z_i + \hat{v}_{Z_i} = \hat{Z}_i \quad (2.10)$$

with X_i being the X component of the object coordinates of point i and so on. All together, the 3d object coordinates of the reference points $\boldsymbol{X}_P$ in the object coordinate system are parameterized in this thesis by

$$\boldsymbol{X}_P = \begin{pmatrix} X_{P,1} & Y_{P,1} & Z_{P,1} & \dots & X_{P,n} & Y_{P,n} & Z_{P,n} \end{pmatrix} \quad (2.11)$$

and the exterior orientation $\boldsymbol{X}_E$ for camera j and image k is parameterized by

$$\boldsymbol{X}_{E,j,k} = \begin{pmatrix} X_{0,j,k} & Y_{0,j,k} & Z_{0,j,k} & \theta_{0,j,k} & \theta_{1,j,k} & \theta_{2,j,k} & [\theta_{3,j,k}] \end{pmatrix} \quad (2.12)$$

with X_0, Y_0 and Z_0 describing the position of the projection center in the object coordinate system and θ_p being one of the 3d rotation parameters describing the rotation from the object coordinate system into the camera coordinate system. Hereby, $p = 0...2$ for Euler angle representation or axis-angle representation and $p = 0...3$ for quaternion representation ([] indicates optional parameters). $R_{1,1,j,k}$ etc. represent the elements of the 3x3 rotation matrix that can be obtained from other rotation representations and vice versa. It should be acknowledged that parameterizing 3d rotations is faced with some challenges [e.g. Albl & Pajdla, 2014]. Using Euler angles is faced with the risk of singularities. Using quaternions requires an additional parameter to fully represent a rotation, i.e. 4 instead of 3; same for using rotation matrices with 9 parameters. Note, these parameters are not fully independent from each other. $f_{x,j}, f_{y,j}, c_{x,j}, c_{y,j}$ and the distortion parameters used for Δx and Δy define the interior orientation $\boldsymbol{X}_I$ for each camera j according to Subsection 2.1.4. If applicable for the prevalent camera system, the relative orientation $\boldsymbol{X}_R$ between two cameras is parameterized the same way as $\boldsymbol{X}_E$. Note that there are slightly different definitions of the collinearity equations in the literature, for example with regard to the order of the rotation matrix elements, with regard to the sign of the focal length or with regard to the symbols used for the orientation parameters [Förstner & Wrobel, 2016; Luhmann et al., 2016]. Note that superscripts (e.g. o denoting the object coordinate system, cf. Subsection 4.2.6) are sometimes omitted for the sake of readability.

2.2.4 Datum definition

Depending on the geometry of the adjustment problem and the given observations, there might be a datum deficiency, which could make explicit datum definition necessary. From the point of view of numerical mathematics, datum definition allows to obtain a unique solution during optimization. Otherwise, the Jacobian matrix $\boldsymbol{A}$, resulting from linearization of the functional model and containing the partial derivatives of all function equations with regard to the unknown parameters, would be singular, and as consequence the normal equation matrix $\boldsymbol{N}$, derived from the Jacobian matrix and necessary for solving the adjustment problem, could not be inverted and so no solution be provided. From the point of view of network geometry, datum definition means introducing additional information resolving the datum deficiency, i.e. the ambiguities in the translation, rotation and scale of a network of points with regard to a higher-level coordinate system. The type of observations and the dimensionality of the network determines which ambiguities need to be resolved. Common approaches for datum definition in photogrammetry are either using free adjustment or defining some points as fixed datum points [Luhmann et al., 2013]. The later can be modeled either as constants with error-free coordinates or as observations with a realistic observation weight. Advantageous of free adjustment is that no undesired constraints on the inner geometry of the network are imposed: If otherwise the number of fixed datum point coordinates is higher than the datum deficiency, such undesired constraints may occur. Furthermore, according to Luhmann et al. [2013], free adjustment provides optimal precision compared to unconstrained or overdetermined datum definition using fixed datum points.

Both error-free datum points and datum points with realistic weights can be modeled as additional observations in the functional model of an adjustment. The stochastic model has to be extended for the additional observations by unrealistically high or realistic observation weights, respectively. Alternatively, error-free datum points can be modeled by removing their object coordinates from the set of unknown parameters. Free adjustment can be modeled by adding constraining functions to the functional model (cf. Subsection 2.2.1), which are derived from the condition equations $\boldsymbol{H}^T \Delta\hat{\boldsymbol{x}} = \boldsymbol{0}$. Hereby, $\boldsymbol{H}$ is a constraint matrix established by partial derivatives of the condition equations with respect to the datum parameters [Förstner & Wrobel, 2016] and $\boldsymbol{\Delta x}$ describes the update of unknown parameters estimated in one optimization iteration, i.e. the interior, if applicable relative, and exterior orientation parameters and the object point coordinates in a bundle adjustment. Thus, $\boldsymbol{H}$ is defined for bundle adjustment with a mono camera as

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial \Delta \boldsymbol{X}_{E,1,1}}{\partial \boldsymbol{T}} & \frac{\partial \Delta \boldsymbol{X}_{E,1,1}}{\partial \boldsymbol{S}} & \frac{\partial \Delta \boldsymbol{X}_{E,1,1}}{\partial \mu} \\ \vdots & \vdots & \vdots \\ \frac{\partial \Delta \boldsymbol{X}_{I,1}}{\partial \boldsymbol{T}} & \frac{\partial \Delta \boldsymbol{X}_{I,1}}{\partial \boldsymbol{S}} & \frac{\partial \Delta \boldsymbol{X}_{I,1}}{\partial \mu} \\ \vdots & \vdots & \vdots \\ \frac{\partial \Delta \boldsymbol{X}_{P,1}}{\partial \boldsymbol{T}} & \frac{\partial \Delta \boldsymbol{X}_{P,1}}{\partial \boldsymbol{S}} & \frac{\partial \Delta \boldsymbol{X}_{P,1}}{\partial \mu} \\ \vdots & \vdots & \vdots \end{bmatrix} \tag{2.13}$$

with $\boldsymbol{T}$ covering the three 3d translation parameters, $\boldsymbol{S}$ covering the three rotation parameters and μ covering the scale parameter of a similarity transformation. $\boldsymbol{X}_P$, $\boldsymbol{X}_I$ and $\boldsymbol{X}_E$ are defined as in Subsections 2.1.4 and 2.2.3. As proposed by Polic et al. [2018] for large-scale camera calibration, the condition equations describe differences of the estimated orientation parameter values and object coordinates before and after applying a similarity transformation that links the network to the higher-level coordinate system and so realizes the datum definition (denoted by left superscript ${}^t(...)$). By setting the right side of the condition equations to zero, it is ensured that the translation, rotation and scale change applied to the network points are in total zero.

Exemplarily, the mentioned differences for the three position parameters $\boldsymbol{X}_{0,1,1}$ of the exterior orientation $\boldsymbol{X}_{E,1,1}$ for camera 1 and image 1 are defined as

$$\Delta\hat{\boldsymbol{X}}_{0,1,1} = \hat{\boldsymbol{X}}_{0,1,1} - {}^{t}\hat{\boldsymbol{X}}_{0,1,1}(\boldsymbol{q_S}) = \hat{\boldsymbol{X}}_{0,1,1} - (\mu\boldsymbol{R}(\boldsymbol{S})\hat{\boldsymbol{X}}_{0,1,1} + \boldsymbol{T}) \tag{2.14}$$

and the differences for the object coordinates $\boldsymbol{X}_{P,1}$ of point 1 as

$$\Delta\hat{\boldsymbol{X}}_{P,1} = \hat{\boldsymbol{X}}_{P,1} - {}^{t}\hat{\boldsymbol{X}}_{P,1}(\boldsymbol{q_S}) = \hat{\boldsymbol{X}}_{P,1} - (\mu\boldsymbol{R}(\boldsymbol{S})\hat{\boldsymbol{X}}_{P,1} + \boldsymbol{T}) \tag{2.15}$$

with $\boldsymbol{q_S} = [\boldsymbol{T}, \boldsymbol{S}, \mu]$ and $\boldsymbol{R}$ being the rotation matrix corresponding to $\boldsymbol{S}$. As the interior orientation parameters are not affected by the described similarity transform, $\Delta\hat{\boldsymbol{X}}_{I,1}$ etc. are zero, and their derivatives are zero as well. According to Polic et al. [2018], $\Delta\hat{\boldsymbol{\theta}}_{1,1}$ etc. addressing the rotation angles of the exterior orientation is not calculated analogue to Equations 2.14 and 2.15. Instead, its derivatives are obtained in a complex multi-step process from parts of the Jacobian matrix $\boldsymbol{A}$. As the method proposed by Polic et al. [2018] supports mono cameras only, it needs to be extended in order to support the stereo camera system that is used to investigate the windshield effects.

2.3 Computer vision on road scene images

For automotive applications, various computer vision tasks are performed on road scene images, wherefore nowadays typically deep learning is employed. In the following, computer vision tasks that are relevant for the proposed camera calibration methods are introduced.

2.3.1 Detection and segmentation

It is state of the art to evaluate road scene images taken with on-board cameras in vehicles by deep learning methods for various computer vision tasks to get a better scene understanding. Relevant tasks for this thesis are (i) object detection, (ii) semantic segmentation, (iii) instance segmentation and panoptic segmentation, (iv) depth estimation and (v) edge or boundary detection. Object detection [Girshick et al., 2014; Liu et al., 2016; Lin et al., 2017; Redmon & Farhadi, 2017; Ren et al., 2017] aims at finding the position of individual objects belonging to a certain class in images. Detected objects are typically marked by enclosing rectangles, which means that no pixel-level object boundaries are obtained. Earlier methods for object detection typically divide an image into smaller sub-images, which are then checked with high effort one after the other whether they contain desired objects [e.g. Sermanet & LeCun, 2011; Houben et al., 2013; Benenson et al., 2015] (sliding window approach). With such an approach the same object may be detected multiple times in nearby sub-images, which needs to be resolved e.g. by post-processing [Hanel & Stilla, 2018, 2019a]. More recent methods evaluate an entire road scene image in one step [e.g. Zhu et al., 2016; Janai et al., 2017] and so provide more consistent detections. Nevertheless, neither pixel-accurate object boundaries nor scene knowledge for the entire image are obtained, which both is relevant for the proposed methods. In contrast, semantic segmentation [Ronneberger et al., 2015; Chen et al., 2016; Shelhamer et al., 2017; Badrinarayanan et al., 2017; Chen et al., 2018b] aims at predicting a semantic image providing pixel-wise information about the semantic class of objects shown in a given RGB image. Typically, the semantic class is determined for every pixel [e.g. Chen et al., 2018b]. The resulting semantic image consists of multiple segments, each belonging to one semantic class. Each segment can contain even more than one individual object of this class, which may be a problem for certain applications [e.g. Liu et al., 2018]. Ideally, the boundaries of the segments match with object boundaries in the RGB image. For semantic segmentation of road scene images, the classes represent common road scene objects like vegetation, vehicle, building or road, whereby often the class definition from Cordts et al. [2016]

is applied. Relevant in particular for automotive applications, Vertens et al. [2017] have proposed to predict the motion status, i.e. whether an object is moving or standing, in addition to semantic classes. Note that besides the image-based semantic segmentation addressed in this paragraph, semantic segmentation could be performed on point clouds as well [Brostow et al., 2008; Charles et al., 2017; Huang et al., 2019]. As a similar computer vision task, instance segmentation can be seen as combination of object detection and semantic segmentation, as it targets at pixel-level segments belonging to individual objects of specific classes [Hariharan et al., 2014; He et al., 2017; Liu et al., 2018]. By instance segmentation, typically only image parts containing objects of desired classes are considered in the resulting semantic images and no semantic information is obtained for other image parts. As consequence, especially non-countable semantic classes like *road, sidewalk* that are relevant for automotive applications might be neglected. For these reasons, instance segmentation is not considered for the proposed methods. Recently, the topic of panoptic segmentation as merge between semantic segmentation and instance segmentation arose [Mohan & Valada, 2021], whereby typically fully-covered semantic images with segments containing individual objects instead of object classes are obtained. Though it would be an alternative to semantic segmentation that is worth to consider, panoptic segmentation is only used for one experimental variation, as most experiments have been performed before its first publication. As next task, single image depth estimation aims at providing disparity and depth maps for given RGB images [Eigen et al., 2014; Garg et al., 2016; Laina et al., 2016]. By depth maps, for each pixel the metric distance of the shown object from the camera is obtained, which is used by one of the proposed methods. While e.g. with stereo cameras or RGB-D cameras, depth maps are obtained by an geometry-based approach, typically machine learning is employed to retrieve them from single images [Mertan et al., 2022]. The last task, edge detection, also called boundary detection, aims at deriving certain types of edges [Marr & Hildreth, 1980; Bertasius et al., 2015; Liu et al., 2019]. Such edge types may be low-level intensity changes in images till high-level object boundaries [Deng et al., 2018]. For one of the calibration methods proposed in this thesis, detecting edges that are object boundaries is required (cf. Chapter 5), wherefore a boundary detection method that utilizes semantic information by deep learning to find the boundaries of objects is used [Yu et al., 2017]. In contrast, classic edge detectors like the Canny edge detector [Canny, 1986] do not employ semantic information and so they seem less applicable to get specifically the boundaries of objects.

2.3.2 3d reconstruction and camera localization

A 3d reconstruction of the environment or camera localization can be obtained by approaches like (i) visual odometry (VO), (ii) multi view stereo (MVS), (iii) visual simultaneous localization and mapping (SLAM) or (iv) structure-from-motion (SfM) using data from different kinds of sensors, like monocular cameras [Engel et al., 2014; Mur-Artal et al., 2015], stereo cameras [Wang et al., 2017], RGB-D cameras [Henry et al., 2012; Nießner et al., 2013], or LiDAR [Jiang et al., 2016; Graeter et al., 2018], for example. Visual odometry aims at camera localization only and does not create an environment reconstruction at all [Guerrero et al., 2005; Nister et al., 2006]. In contrast, MVS aims at obtaining an environment reconstruction for given camera orientations [Labatut et al., 2007; Furukawa & Hernández, 2015]. Due to their objectives, these two kinds of approaches are not considered as relevant for camera calibration within this thesis. Visual SLAM can realize consistency between 3d reconstruction and camera orientations by constraining the camera trajectories, for example by loop closures [Scaramuzza & Fraundorfer, 2011; Yousif et al., 2015], but aims at real-time performance. In contrast, SfM typically doesn't aim at real-time performance, therefore allowing computationally expensive offline optimization (i.e. bundle adjustment) to obtain global consistency in the 3d reconstruction. Therefore, SfM is seen as more relevant for the self-calibration methods proposed in this thesis. These four kinds of approaches

can be either indirect methods [Schönberger & Frahm, 2016; Mur-Artal & Tardós, 2017] relying on feature descriptors like SIFT [Lowe, 1999], SURF [Bay et al., 2006] or ORB [Rublee et al., 2011], or direct methods [Wang et al., 2017; Engel et al., 2018] comparing intensities between different image patches. For the later ones, problems have been reported for auto exposure cameras and in the case of vignetting [Bergmann et al., 2018], which both could play a role for on-board cameras. Furthermore, the performance of direct methods may suffer in the case of large motions between consecutive images [Younes et al., 2019], which could apply for images recorded at high vehicle velocities. Therefore, indirect methods are used in this thesis. Comparing different descriptors for indirect methods, Tareen & Saleem [2018] have reported in their comparative analysis of matching performance that SIFT has shown the best accuracy despite its age. As consequence, in this thesis SIFT is used for feature extraction for camera calibration by semantic structure-from-motion.

3 State of the art

This chapter covers the state of the art in automotive camera calibration that is relevant for the proposed methods. At first automotive test field calibration is addressed (Section 3.1), covering the motivation to use test field calibration to investigate the influence of the windshield and also covering two challenges with regard to the selection of the test fields. Furthermore, different approaches to cope with the windshield effects are discussed. Then automotive self-calibration is addressed (Section 3.2), whereby different types of reference information are described that can be derived from road scene environments. In particular, the benefits and challenges of using traffic signs and image features to derive the reference points for automotive self-calibration are addressed. Finally, two important aspects for self-calibration with reference points from image features are discussed: The approaches to obtain semantic scene knowledge and their integration into the 3d reconstruction methods used for self-camera calibration.

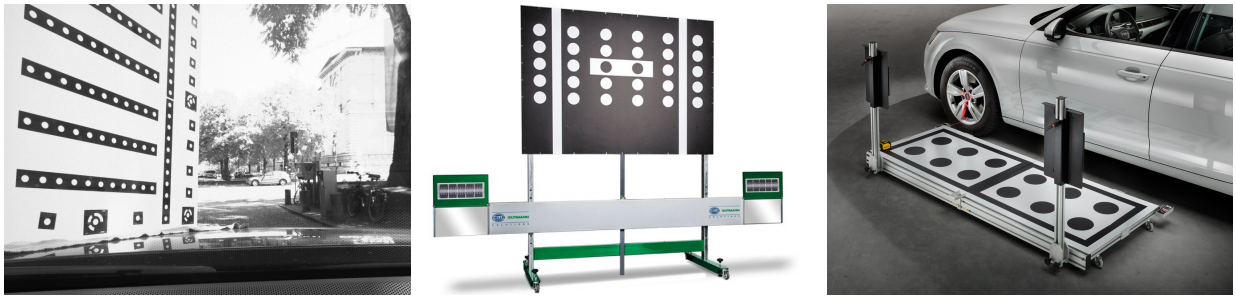
3.1 Automotive test field calibration

First, the selection of the test fields for automotive camera calibration and, second, the windshield refraction are addressed.

3.1.1 Test fields in automotive camera calibration

Automotive self-calibration depends on the availability of appropriate road scene-specific reference information (e.g. lane markings) at the desired calibration locations. The estimated values of the orientation parameters are more likely to be scene-dependent, highly correlated and only valid for the part of the image area that is covered by reference points, like the lower image half for lane markings [Fraser, 1997, 2013]. Thus, automotive self-calibration often requires additional algorithmic steps, e.g. to remove outliers by point filtering, for a reliable calibration [Dang et al., 2009; Lasaruk & Hachfeld, 2019]. In contrast, automotive test field calibration is robust to varying imaging conditions and can be performed in a well-controlled environment [Geiger et al., 2012; Rosebrock & Wahl, 2012], so that a high quality of the calibration results and comparability between different calibration iterations can be expected. For these reasons, the investigation of the windshield influence (see Chapter 4) is carried out by test field calibration.

Though, two aspects need to be considered when using test field calibration for this investigation. First, space restrictions, either induced by vehicle mass-production processes or by limitations in research facilities, could constrain the position and orientation of test fields (Figure 3.1) used for calibration of vehicle cameras [Scheller et al., 2007; Bodis-Szomoru et al., 2008]. These constraints can occur in a way so that no sufficient coverage of all six degrees of freedom (DOF) may be achieved [Schöller et al., 2019; Muhovic & Pers, 2020], which could impose a negative influence on camera calibration [Hastedt et al., 2016]. Stereo camera calibration is affected by space restrictions in particular, as even more space is required to place test fields covering both fields of view adequately [Kruger et al., 2004; Bodis-Szomoru et al., 2008; Lasaruk



(a) (b) (c)

Figure 3.1: Test fields for automotive camera calibration. a) Classic 2d photogrammetric test field, b) automotive-specific test field for forward-looking cameras [Hella Gutmann Solutions GmbH, 2019], c) automotive-specific test field for surround-view cameras [Texa, 2022].

& Hachfeld, 2019]. Even for test field calibration with known precise object coordinates of the reference points, an appropriate imaging geometry has to be ensured so that good precision and low correlations between the estimated parameters can be expected [Mitishita et al., 2009; Fraser, 2013]. Therefore, the problem arises to select a test field for the investigation of the windshield influence that allows a sufficient coverage of all DOFs.

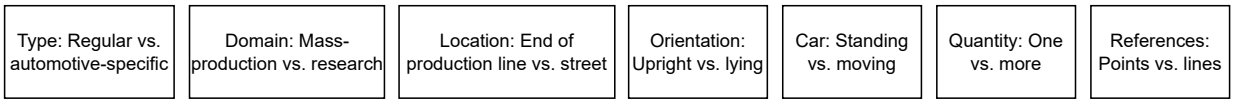


Figure 3.2: Common characteristics of test fields in automotive camera calibration.

Second, a plenty of automotive-specific test fields exist for different types of vehicle cameras [Marita et al., 2006; Scheller et al., 2007; Friel et al., 2012; Rosebrock & Wahl, 2012; Pliefke, 2013; Hanel et al., 2016; Hella Gutmann Solutions GmbH, 2016; Thatcham Research and ADAS Repair Group, 2016; Texa, 2017; Robert Bosch GmbH, 2018] (Figure 3.2): While forward-looking cameras are typically calibrated with upright standing test fields, test fields lying on the ground are used for surround-view systems mainly [Geiger et al., 2012; Pliefke, 2013]. Common automotive-specific test fields may provide only a small number of reference points [Hella Gutmann Solutions GmbH, 2016; Texa, 2017]. As in addition to a sufficient coverage of the DOFs, obviously a large number of reference points is desired for a good imaging geometry for camera calibration, also such aspects as mentioned have to be taken into account when selecting the number and type of test fields for the proposed method.

3.1.2 Windshield refraction in automotive camera calibration

One objective of this thesis is to investigate the influence of the vehicle windshield on automotive test field calibration. The presence of such an influence has been addressed already in previous work: Zou & Li [2015] state that windshield refraction is important for calibration of cameras inside the car that observe test fields outside. Lasaruk & Neralla [2018] suggest that compensation models for the windshield effects obtained for one windshield can be applied to other windshields of the same model, while Dlugosz et al. [2019] believe that calibration is necessary for each individual camera due to differences in the effects. The refractive effects can either be compensated implicitly by using standard camera models or explicitly by using an extended camera model [Kahmen et al., 2020]. In this thesis, an implicit solution is used so that potential effects of the windshield can become visible as differences in the calibration results compared to calibration without windshield. As stated by Verbiest et al. [2020], only few work has been published on the assessment of refractive

effects: They have shown that differences in the values of the interior orientation parameters due to windshield effects are larger for implicit than for explicit compensation. But to the best of the author's knowledge, none of the previous works has focused on assessing differences in the uncertainties (standard deviations, correlations) of the camera orientation parameters due to windshield effects. Therefore, special emphasis of the investigations carried out with the proposed method for test field calibration is put on the uncertainties. A stereo camera system is used for these investigations, as they are especially sensitive to calibration errors that could result from windshield refraction, as already outlined in the previous chapters.

3.2 Automotive self-calibration

To ensure valid calibration parameters over vehicle lifetime, self-calibration is interesting as test field calibration done while driving on the road is typically not realistic: The effort to provide test fields at the desired calibration locations and time points on public roads would be infeasibly high [Pagel & Willersinn, 2011; Schöller et al., 2019]. Furthermore, special test fields for use on the road, like markers placed on the street surface [Marita et al., 2006; Bodis-Szomoru et al., 2008] or on the vehicle hood [Broggi et al., 2001], may be applied in research settings only [Gil et al., 2018b], same as returning to special calibration facilities from time to time [Dang et al., 2009; Hold et al., 2009; Friel et al., 2012]. With regard to either geometric, temporal or stereo constraints used for automotive self-calibration, the employed reference information obviously originates from the ego-car or the road scene environment (Figure 3.3) as described in the following.

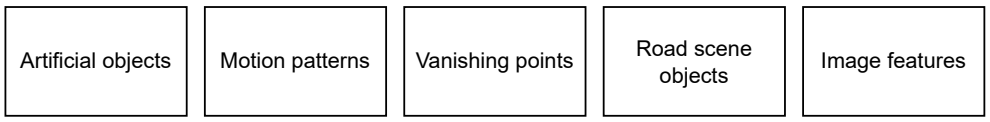


Figure 3.3: Common types of reference information for automotive camera self-calibration.

3.2.1 Artificial objects and ego-car motion

Artificial objects, like special patterns of road markings, can not only be utilized for test field calibration, but also for automotive self-calibration [Tan et al., 2011]. Same as for test field calibration, the most prominent drawback is the required high manual effort to handle them [Bhardwaj et al., 2018]. This applies also if only parts of the reference information should be provided by artificial objects, for example if metric scale information should be obtained from scale bars placed in the road scene [Scheller et al., 2007; Stein et al., 2010; Heng et al., 2015; Knorr, 2018]. So, especially for mass-produced vehicles, such approaches can't be seen as feasible. Besides artificial objects, also ego-car motion can be used to derive reference information, either by having knowledge about the motion or by demanding certain motion patterns to be performed; for example, translation-only motion, i.e. driving straight, or doing a cornering maneuver [Miksch et al., 2010; Houben, 2014; Paula et al., 2014; Lasaruk & Hachfeld, 2019]. Though, demanding certain patterns seems also not to be feasible, as driving specified maneuvers may be challenging for regular drivers without special instructions or training. In contrast, as modern vehicles are equipped with many additional sensors, using wheel odometry [Heng et al., 2013; Okouneva, 2017], GPS or IMU [Gopaul et al., 2016; Mueller & Wuensche, 2017; Borgmann et al., 2018; Hanel & Stilla, 2019b], or radar [Muhovic & Pers, 2020] may be the more feasible alternative to provide reference information that originates from the motion of the ego-car. Therefore, vehicle positions obtained from GPS are used in the proposed method for camera calibration by semantic structure-from-motion to get metric scale information, which arises the need to appropriately

integrate these positions into the structure-from-motion algorithm. As additional requirement, the positions should match realistic vehicle motion patterns to get the best-possible scale.

3.2.2 Vanishing points

Vanishing points for calibration require orthogonal lines to model constraints for self-calibration [Tan et al., 2011]. Such orthogonal lines may originate from different kinds of reference objects, like a set of orthogonal lane markings [Alvarez et al., 2014], like buildings following the so-called Manhattan directions [Lu et al., 2013], like vertically aligned light poles [Lu et al., 2013; Alvarez et al., 2014] or like upright standing pedestrians [Corres et al., 2016]. Vanishing points may also be derived from orthogonal moving directions of cars, e.g. at intersections [Alvarez et al., 2014]. Often not a single kind of objects is sufficient for full calibration, so e.g. two orthogonal lines from lane markings need to be combined with a third line from vertical poles. Furthermore, cars or pedestrians may occlude important lines needed. As consequence, relying on a set of orthogonal lines limits calibration to scenes where the desired reference objects are available. Additionally, inappropriate lines, e.g. along short edges like small window tiles, could lead to inaccurate vanishing points or make it difficult to clearly identify orthogonal lines [Zhou et al., 2017]. Thus, the complexity of the calibration algorithm increases and the applicability to unknown road scenes decreases, as image processing needs to be tailored to the expected appearance of the desired objects [Chang & Tsai, 2012].

3.2.3 Reference information from characteristics of road scene objects

Reference points for calibration may be derived from characteristic points at stationary road scene objects like corner points of road markings [Catala-Prat et al., 2006; Ribeiro et al., 2006; Hold et al., 2009] or traffic signs [Lamprecht et al., 2007; Hanel & Stilla, 2018; Lasaruk & Hachfeld, 2019]. Under special conditions, reference points can be derived even from moving objects, like from tail lights of vehicles [Bhardwaj et al., 2018]. Also more complex reference information than points may be utilized, like the characteristic shape of the road surface in the upcoming driveway [Musleh et al., 2014]. For example, assuming a planar shape of the road surface allows to define constraints for camera calibration based on depth values estimated with a stereo camera system [Garcia, 2017; Muhovic & Pers, 2020]. Alternatively, Catala-Prat et al. [2006] suggest to iteratively update the exterior orientation parameters until the characteristic parallelity of a pair of road markings, transformed from perspective to orthographic projection, is fulfilled. Using such characteristics of road scene objects for camera calibration is faced with several challenges. Road markings or the road surface are typically shown in lower parts of a road scene image, wherefore calibration will be valid only for these image parts with reference information [Luhmann et al., 2006; Hanel & Stilla, 2018]. Furthermore, similar as for vanishing points, calibration utilizing certain types of objects is restricted to roads where these objects are present [Bertozzi et al., 2010; Pflug et al., 2013] and not occluded by other objects, which can easily happen for objects on or low above the ground [Musleh et al., 2014; Häne et al., 2017]. Additionally, assumptions on object-specific characteristics like straight and parallel road markings or a flat ground plane [Fung et al., 2003; Catala-Prat et al., 2006; Paone et al., 2019] could be problematic for calibration, if the real objects deviate from these assumptions. These challenges can be overcome by traffic signs (i) that are typically shown in upper image parts to get calibration results that are, in combination with e.g. road markings, valid for the entire image, (ii) that are frequently present on public roads, (iii) that impose no further requirements on the scene, like a flat road, and (iv) that are standardized by official regulations so that their shape and metric size are known what allows to develop algorithms to identify characteristic points by image processing in order to use them as reference points. While in their previous work Lamprecht et al. [2007] only theoretically define

several arbitrary points at traffic signs as reference points for calibration, simulate their object coordinates for evaluation and assume the vehicle speed to be known, require Lasaruk & Hachfeld [2019] the car to drive a cornering maneuver and just mention traffic signs as one of several potential stationary objects that may be used for calibration. Hence, to the best knowledge of the author, the challenge remains to develop an algorithm for camera self-calibration with traffic signs covering the entire workflow from a sequence of real images until estimation of the camera orientation parameters.

3.2.4 Reference points from image features

Reference points for calibration can be also obtained from arbitrary road scene objects using image feature detectors and descriptors [Ruland et al., 2010; Cannelle et al., 2012; Livyatan & Berberian, 2017; Okouneva, 2017; Hanel & Stilla, 2019b], wherefore no characteristics of the objects have to be known or identified. Both histogram-based, like SIFT [Lowe, 1999], and binary descriptors, like BRIEF [Calonder et al., 2012], as well as optical flow [Pflug et al., 2013] have been utilized in previous work to get reference points for automotive self-calibration [Mentzer et al., 2017]. To establish the constraints required for self-calibration, correspondences between features points have to be obtained, either between points in images of different cameras in a vehicle multi-camera system taken at the same time or between points in images of a vehicle mono camera taken at different times [Ruland et al., 2010; Hansen et al., 2012; Heng et al., 2013; Winner et al., 2015; Pekkucuksen & Batur, 2018]. Camera calibration with reference points from image features can be beneficial in comparison to the previously described types of reference information, as it can be expected that (i) a larger number of reference points and (ii) a better distribution of them in the scene can be obtained leading to more reliable calibration results, as well as (iii) that larger parts of the image are covered by reference points for which the estimated orientation parameters are valid. Though, reliable vehicle camera self-calibration with reference points from image features is faced with two key challenges with regard to (i) the static scene assumption [Lasaruk & Hachfeld, 2019] and (ii) the textures of the road scene objects that need to be resolved. First, intentional and unintentional movement of objects like vehicles or pedestrians, or the movement of trees in the wind violates the static scene assumption and will alter the scene geometry between images taken at different points in time [Dang et al., 2009; Hanel & Stilla, 2019b]. Furthermore, reference points on moving objects can be easily occluded at some time points [Musleh et al., 2014], which arises the problem to detect moving objects in order to avoid reference points on them for automotive self-calibration. Second, commonly the objects need to have suitable textures for feature extraction and matching [Pflug et al., 2013]. Lack of sufficient texture, e.g. poor-textured objects like sky, repetitive surfaces like at tarmac, reflecting surfaces like at building windows or metallic car paint [Ruland et al., 2010] can lead to invalid reference points. Hence, the problem arises to avoid reference points at such textures.

3.2.5 Semantic 3d reconstruction methods for automotive camera calibration

Automotive self-calibration relying on image features can be realized by 3d reconstruction and localization methods like visual SLAM [e.g. Mur-Artal & Tardós, 2017] or structure-from-motion [e.g. Schönberger & Frahm, 2016]. Thereby, the calibration parameters are estimated as side product besides creating a 3d reconstruction of the road scene [Heng et al., 2014]. Beneficially, even sparse 3d reconstructions provide - depending on the number of images - often several thousand reference points, which has been reported to be an important factor for camera calibration [Stamatopoulos & Fraser, 2014]. Additionally, the 3d reconstruction enables calibration of multiple cameras in the ego-car without overlapping fields of view and even calibration of cameras in other cars [Leite et al., 2008; Heng et al., 2014]. Obviously, the mentioned challenges with

moving objects and inappropriate textures (cf. Subsection 3.2.4) may be overcome by generic outlier removal steps, e.g. based on RANSAC, that are often part of 3d reconstruction methods [e.g. Schönberger & Frahm, 2016]. Though, generic outlier removal is not tailored to the mentioned kinds of road scene objects and so bears the risk to either consider a too small or too large number of reference points as outliers, which both is negative for camera calibration. Therefore, it is desired to apply a specific outlier removal tailored exactly to the problematic road scene objects causing the mentioned challenges. In order to know whether such objects are present in the scene and where they are shown in an image, scene knowledge needs to be obtained. For 3d reconstruction and localization methods, scene knowledge may be obtained by (i) semantic object databases, (ii) instance segmentation, (iii) object detection or (iv) semantic segmentation. First, such pre-built semantic object databases that are linked to the images used for 3d reconstruction based on matched image features [Civera et al., 2011] may not be available for the large variety of types and appearances of road scene objects, wherefore this approach is not considered as suitable. Second, instance segmentation [Rünz & Agapito, 2017; Runz et al., 2018; Barsan et al., 2018; Wang et al., 2018] and third, image-based object detection [Bao & Savarese, 2011; Sünderhauf et al., 2017; Qi et al., 2018] provide scene knowledge typically for certain image parts and for a small number of semantic classes only (cf. Section 2.3). As the image points of the reference points for calibration may be located in all image parts and as enclosing rectangles resulting from object detection do not match the object boundaries exactly (cf. Section 2.3 as well), also these two approaches are not considered as suitable. Forth, as semantic segmentation typically provides scene knowledge for the entire image [Stueckler et al., 2012; Yu et al., 2018], image points in all image parts can be handled. Furthermore, semantic segmentation typically supports a larger set of semantic classes covering typical road scene objects [Cordts et al., 2016]. For these two reasons, semantic segmentation appears to be most suitable to obtain scene knowledge.

There are also several approaches how the scene knowledge from semantic segmentation can be integrated into a 3d reconstruction and localization method. First, semantic segmentation can be used for localization within an existing 3d reconstruction [Hirzer et al., 2017; Schönberger et al., 2018] and, second, semantic segmentation can enrich a 3d reconstruction in post-processing by assigning semantic information to the 3d points [Li & Belaroussi, 2016; Mahe et al., 2018; Runz et al., 2018]. As it is desirable to use the scene knowledge already while the 3d reconstruction is incrementally created to overcome the described challenges, such post-processing approaches are not applicable. Third, scene knowledge can be also integrated during 3d reconstruction, e.g. for feature tracking [Murali et al., 2017]. Though, these authors work with gray value images only and aim at real-time performance, which is not desirable for camera calibration as it may have negative influence on the quality of the calibration results. Furthermore, they assume sensors that have been calibrated beforehand. More related to the mentioned challenges, Wang et al. [2018] and Yu et al. [2018] use semantic segmentation to obtain knowledge on the presence of moving objects in images in order to remove outliers to make 3d reconstruction more robust. While their works underline the potential of moving objects to cause problems in 3d reconstruction, they focus on moving objects only, but not on road scene objects with inappropriate textures. Furthermore, they consider only a limited set of semantic classes that typically does not cover the entire area of a road scene image, so that problematic points in uncovered parts of the image are not handled. Last, Kaneko et al. [2018] propose to exclude image parts from feature extraction based on scene knowledge obtained by image-based semantic segmentation. Though, they work with visual SLAM, which typically aims at real-time capability, instead of structure-from-motion, which is considered as better suitable for camera calibration (cf. Chapter 2). Furthermore, their objective is to improve the mapping and localization performance, but not the results of camera calibration. Finally, they test their method on synthetic images only, which leaves it open to show the applicability on real images. Concluding from these findings, it remains unsolved yet to show

the potential of semantic knowledge in a 3d reconstruction method to improve self-calibration and to apply the developed workflow on real-world images. Furthermore, it is unsolved yet to integrate the semantic knowledge in other steps of the 3d reconstruction method than feature extraction, for example during feature matching.

4 Camera calibration with test fields through a vehicle windshield

In this chapter, a method for stereo camera calibration with test fields is described that is used to investigate the influence of a vehicle windshield on calibration. The investigation is carried out as comparative analysis between two setups, one time with and the other time without a windshield in the optical path between the cameras and test fields during calibration. First, the intended calibration setup comprising a virtual 3d test field and a stereo camera system is introduced (Section 4.1). Then, a detailed description of the calibration workflow from image acquisition until estimation of the orientation parameters and their uncertainties follows (Section 4.2). Special emphasis is put on the realization of the virtual 3d test field and on datum definition.

4.1 Calibration setup

The calibration setup comprises a virtual 3d test field that is created by two non-rigid 2d test fields (Subsection 4.1.1), the stereo camera system with industrial cameras with approximately parallel optical axes and a baseline that is typical for the automotive domain (Subsection 4.1.2) and, in one setup, a vehicle.

4.1.1 Virtual 3d test field

This method is intended for a virtual 3d test field consisting of two independent, non-rigidly coupled 2d test fields to provide the reference points for calibration. The virtual 3d test field can be handled easier than a rigid 3d test field and hence, a good coverage of the six degrees of freedom can be achieved without the cumbersome need to move the vehicle carrying the cameras, which is anyways limited to the ground plane. Additionally, the risk of suffering from strong correlations between the estimated interior and exterior orientation parameters is lower than when using a single 2d test field [Luhmann et al., 2016]. Though, the disadvantage of the virtual 3d test field is that the relative orientation between the two 2d test fields changes from one time point to the next due to the missing rigidity. Hence, correspondences between the reference points on the two test fields need to be established for each time point and a joint 3d object coordinate system needs to be determined for calibration, as it will be described in a later section of this chapter.

This method is designed for different types of reference marks on the 2d test fields, therefore benefiting from a lower risk that no reference points can be detected by image processing at all, for example in the case of large distances from the camera or unfavorable scene illumination. One test field should have a dense grid of circular marks, while the other should have a classic checkerboard pattern. Both are cheap, allow easy and robust image processing, provide a sufficient number of reference points and their pixel coordinates can be determined with high accuracy [Geiger

et al., 2012; Lasaruk & Neralla, 2018]. While extracting pixel coordinates from circular marks is always confronted with the problem of ellipse eccentricity [Heikkila & Silven, 1997; Heikkila, 2000], checkerboard patterns provide only a lower density of reference points. Coded and uncoded circular marks as well as even and odd numbers of checkerboard squares, respectively, allow to uniquely identify each reference point. Automotive-specific test fields are not seen as useful, as they typically provide a lower number of reference information than classic photogrammetric test fields (cf. Chapters 2, 3).

4.1.2 Stereo camera system

The method proposed in this chapter is intended for environment-observing, forward-looking stereo cameras, which are chosen for their ability to perform accurate distance measurements in the upcoming driveway of a vehicle, which is in particular relevant for automotive applications (cf. Section 1.1). Monochromatic industrial cameras with a geometric resolution similar to those used in other research related to automotive applications, e.g. for recording the Cityscapes dataset [Cordts et al., 2016; Onsemi, 2017], are suitable in particular. Such industrial cameras have several advantages with regard to the geometric quality of cameras and lenses that allow to avoid certain sources of errors [e.g. Fraser, 2013]: They have fixed focal length and fixed focus, they can be rigidly attached to a stable platform and they typically neither have a complex color filter array in the sensor nor internal image pre-processing algorithms that may alter the raw images and so potentially decrease the quality of the image points of the reference points. Furthermore, the used cameras are capable of simultaneous image acquisition to avoid inconsistencies due to movements of the test fields. The cameras are mounted in a standard stereo configuration for automotive applications, i.e. with parallel optical axes and pixel coordinate axes [Bodis-Szomoru et al., 2008] (Figure 4.1). The field of view is overlapping; this facilitates establishing cross-camera correspondences between reference points on different test fields.

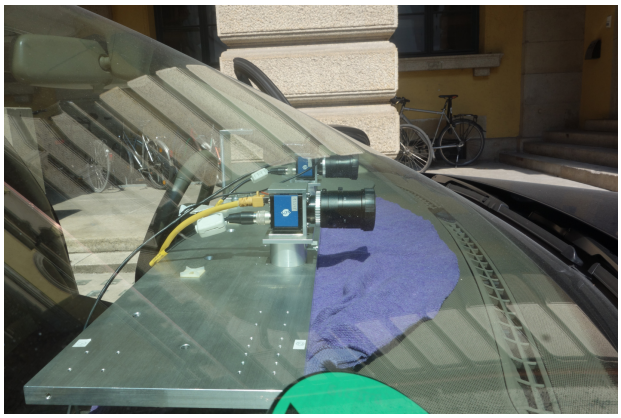


Figure 4.1: Stereo camera system behind the windshield in a car. The cameras are mounted on a rigid metal platform and have approximately parallel optical axes.

4.2 Calibration workflow

The calibration workflow consists of three major steps (red, blue, green areas in Figure 4.2), which are performed for all images from the left and right stereo camera. High-quality object coordinates of the reference points on the test fields are pre-determined once before camera calibration (Subsection 4.2.1). First for each calibration, images sequences are acquired (Subsection 4.2.2). The pixel coordinates of the reference points are extracted from these images and the points are matched across all images (Subsection 4.2.3). The proposed method utilizes reference points from image pairs, consisting of a left and right camera image taken at the same point in time, as well as from independent images, i.e. from images where no reference points could be extracted from the corresponding image of the other camera. Second, the pre-determined object coordinates of the reference points are associated with the pixel coordinates. Special handling is required for uncoded reference marks (Subsection 4.2.4) and for the virtual 3d test field (Subsection 4.2.5).

Third, the final orientation parameters and their uncertainties are estimated by bundle adjustment (Subsection 4.2.6), whereby datum definition is obtained by free adjustment (Subsection 4.2.7).

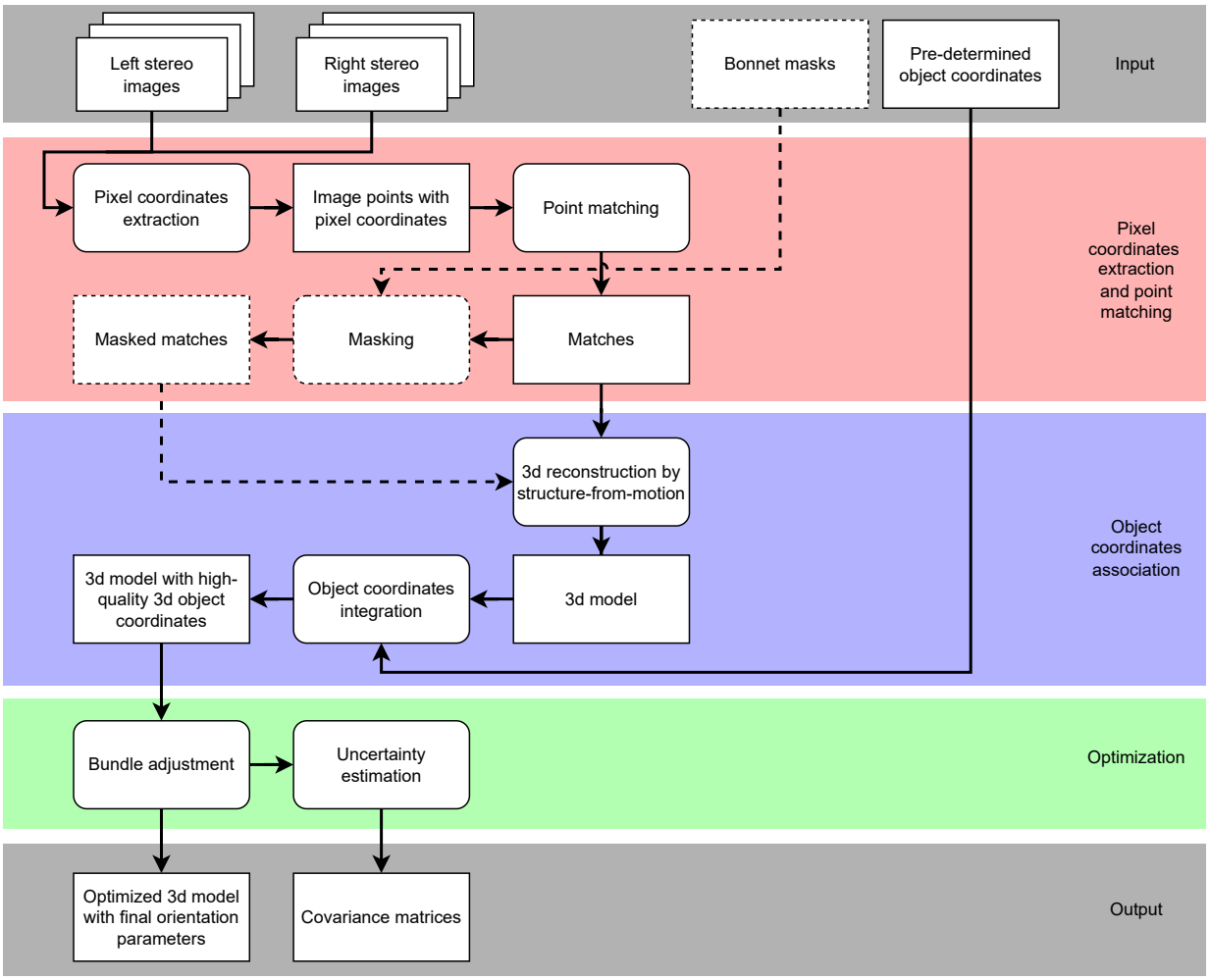


Figure 4.2: Workflow for stereo camera calibration with test fields. Dashed parts are executed for the *without windshield* setup only.

4.2.1 Preliminary steps

The 3d object coordinates of the reference points are pre-determined once before calibration: For the test field with circular marks, a calibrated photogrammetric camera, a high-precision reference cross and reference bar provide a metric-scaled Euclidean object coordinate system in an independent measurement campaign without a windshield. Estimation of the object coordinates is done by bundle adjustment. For the checkerboard test field, the object coordinates are obtained by analytic calculations assuming a grid shape of the checkerboard squares and a known metric size of the squares. Classic camera calibration with a single 2d test field is done separately for each camera to obtain the initial guess for the interior orientation for bundle adjustment.

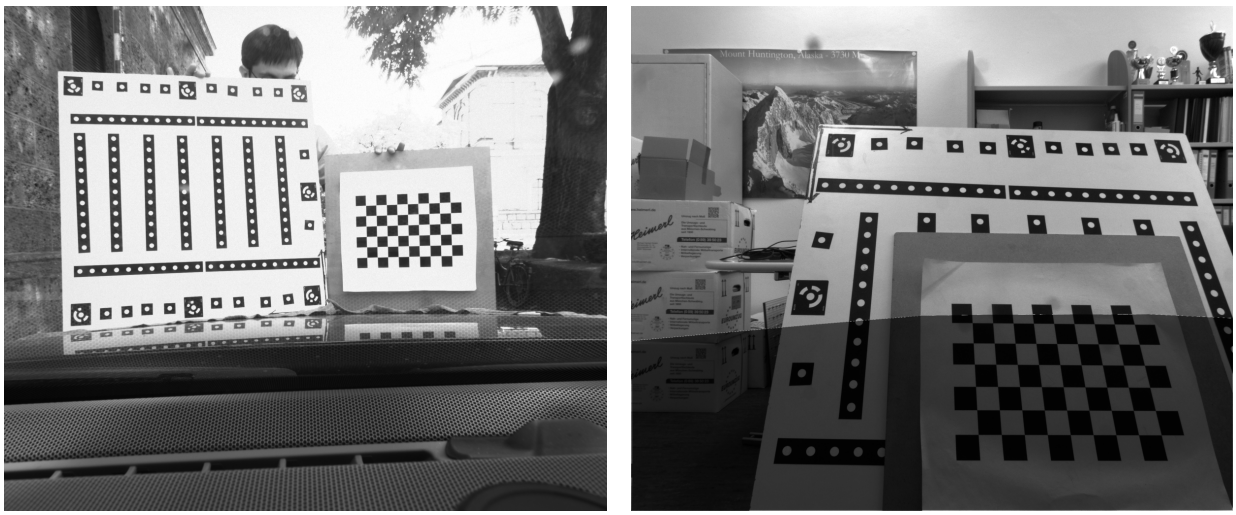
4.2.2 Image acquisition

In the setup with the car windshield as well as in the setup without the windshield, a sufficient number of image pairs is taken with various positions and orientations of the virtual 3d test field relative to the cameras so that the best-possible coverage of the six degrees of freedom can be

achieved. To cope with problematic movements of the hand-held test fields, the images of both cameras are acquired simultaneously. The largest-possible depth range is covered, which means the closest test field position being right in front of the windshield and the most remote position being defined by the ability of the image processing algorithms to extract the reference point coordinates from the images.

4.2.3 Pixel coordinates extraction and point matching

The pixel coordinates of the image points of the reference points are extracted by image processing. For circular marks, this is done by ellipse detection and shape fitting, for the checkerboard pattern this is done by detection of the checkerboard corners, both using standard algorithms. In order to obtain a 3d model of the reference points with initial object coordinates, the image points are matched across cameras and time points. For coded circular marks, the matching degrades to assignment based on the point numbers provided by the code of the reference marks. Points of uncoded circular marks can be matched by point numbers as well. Therefore, unique point numbers are obtained using the known exterior orientations of images that have been determined using the reference points with coded marks. Point matching for the checkerboard test field utilizes unique point numbers that can be assigned if the orientation of the checkerboard pattern in an image is known, which is the case if one grid direction has an even and the other has an odd number of corners.



(a) (b)

Figure 4.3: Example images from the *Stereo image dataset* showing the virtual 3d test field. a) *With windshield* setup: Cameras in the car, b) *without windshield* setup: Cameras in the lab. The bonnet area is masked (half-transparent dark overlay) so that reference point positions are comparable in both setups.

As in the *with windshield* setup, typical for forward-looking automotive cameras, a part of the image always shows the bonnet (cf. Figure 4.3a) and so no points can be found in this part of the image, the reference points in the *without windshield* setup are restricted to the non-bonnet part of the image area to ensure comparability between the two setups. The restriction is implemented by a bonnet mask (dark gray overlay in Figure 4.3b) applied to all images from this setup.

4.2.4 Object coordinates association for uncoded reference marks

In order to achieve the best-possible quality, the pre-determined high-quality object coordinates instead of the initial object coordinates should be used for camera calibration. Therefore, the

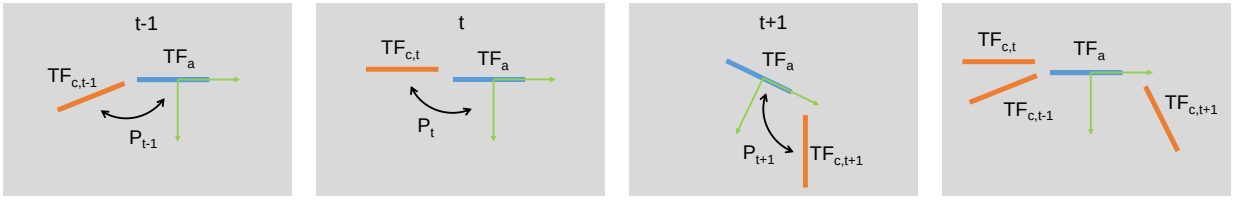
pre-determined object coordinates have to be associated to the pixel coordinates of the reference points. This can be done easily for the reference points from the coded marks, as the same point numbers from the codes are known when the pre-determined object coordinates are obtained (Subsection 4.2.1) and when the pixel coordinates are extracted (Subsection 4.2.3). Likewise, the link can be achieved easily for the reference points from the checkerboard pattern exploiting the even and odd number of checkerboard squares. In contrast, for reference points from uncoded marks, the first set of point numbers that is assigned while obtaining the pre-determined object coordinates is different from the second set of point numbers that is assigned during pixel coordinates extraction and point matching. To solve this problem, a two-folded approach using a 3d model coordinate system (MCS) is employed for reference points from uncoded marks: On the one side, the initial object coordinates of the reference points in the MCS are linked to the extracted pixel coordinates by the second set of point numbers. On the other side, the same object coordinates in the MCS can be linked by geometric transformation to the pre-determined object coordinates that are given in a pre-determined object coordinate system (PCS) with the first set of point numbers.

First, the initial object coordinates of all reference points in the randomly-defined MCS are taken from the 3d model of the reference points obtained by structure-from-motion using the matched image points. Second, the MCS and PCS are associated to each other using the reference points from coded marks that have the same point numbers in both coordinate systems by estimating a 3d Helmert transformation between the MCS and the PCS. Then, third, for all reference points from uncoded marks, the initial object coordinates are projected from the MCS into the PCS using the Helmert transformation and the geometrically closest point in the set of pre-determined object coordinates is determined. With these closest points, the link between the first and second set of point numbers is established and so the pre-determined object coordinates are associated to the pixel coordinates for the uncoded reference marks. Forth and finally, the pre-determined object coordinates are transformed from the PCS into the MCS in order to get a consistent 3d model of the reference points in the MCS. At this stage, the 3d model covers the pixel coordinates and uniquely associated high-quality pre-determined object coordinates of all reference points in the MCS.

4.2.5 Object coordinates association for the virtual test field

In order to allow that reference information from both the test field with circular marks (TF_a) and the checkerboard test field (TF_c) can be used together for calibration, their object coordinates have to be in one common object coordinate system (Figure 4.4d). As the two 2d test fields are only rigid to each other at one time point, but not from one time point to the next (cf. Figure 4.4a-c), the relative orientation P_k between them depends on the time point k. Additionally, as the pre-determined object coordinates are obtained independently (cf. Subsection 4.2.1) and so are given in two independent coordinate systems PCS_a and PCS_c, defining a common object coordinate system and especially incorporating the pre-determined object coordinates as described in the previous subsection is not trivial if both test fields should be used jointly. To overcome this problem, the following approach is employed: On the one hand, for the reference points from TF_a, a single set of reference point numbers that is valid for all time points is defined. On the other hand, for the reference points from the other test field TF_c, a separate set of point numbers is defined for each time point. Thereby, TF_c is treated as there would be actually k independent test fields $TF_{c,k}$, one for each time point k (Figure 4.4d). This causes that the 3d model created by structure-from-motion contains a single set of object coordinates (given in the MCS) for TF_a and separate sets for $TF_{c,k}$. Thus, for TF_a a single Helmert transformation is estimated between PCS_a and MCS, as described in the previous subsection. Additionally, for $TF_{c,k}$ k separate Helmert transformations are estimated between PCS_c and MCS. Having obtained

the transformation parameters, the pre-determined object coordinates are projected into the MCS the same way as described in the previous subsection.



(a) (b) (c) (d)

Figure 4.4: Top-down view on the virtual 3d test field (green, blue) and the object coordinate system (green) at different points in time. a) - c) Real-world view with time-dependently varying relative orientations P_k between the two non-rigid 2d test fields $TF_{c,k}$ (orange) and TF_a (blue) for the time points $k \in \{t-1, t, t+1\}$, d) virtual view with all test fields $TF_{c,k}$ and TF_a in the common model coordinate system used for calibration.

4.2.6 Bundle adjustment

After having associated the object coordinates with the pixel coordinates, the 3d model is optimized by bundle adjustment and hereby the final camera orientation parameters are estimated. Stereo rig constraints are considered in the optimization by one of three *stereo constraints*, which apply different formula and parametrizations for the relative $\boldsymbol{X}_R$ and exterior orientation for the slave camera $\boldsymbol{X}_{E,s}$. Bundle adjustment is selected for optimization due to a variety of advantageous properties for the envisaged investigations (cf. Chapters 2 and 3): Better expected precision, simultaneous estimation of all desired parameters, error minimization with a meaningful geometric constraint and a high redundancy. Typical disadvantages of bundle adjustment like the high computational effort or the need for an initial guess do not play a role as no real-time requirements need to be met in this research.

Optimization is done separately with three camera and distortion models for the sake of comparison (cf. research questions in Section 1.3). Unlike suggested by Maas [2015a] for classic multimedia photogrammetry, the refraction of the optical rays at the windshield is not handled explicitly by extended geometric models. In contrast to the typically flat sheets of glass in multimedia photogrammetry, the complex geometry of a curved vehicle windshield might be highly difficult or even impossible to model accurately. Furthermore, it is an objective of the investigation to show potential effects of the vehicle windshield in the estimated calibration parameter values and their uncertainties.

The stereo constraints are defined as follows:

All stereo constraints

For all three stereo constraints, unknown parameters ($\boldsymbol{X}$) in the bundle adjustment (cf. Subsection 2.2) are the interior orientation and distortion parameters as defined by the camera model (cf. Subsection 2.1.4) and the object coordinates of the reference points. The functional equations are as follows:

$$x'_{i,j,k} + \hat{v}_{x'_{i,j,k}} = f_{x,i,j,k}(\hat{\boldsymbol{X}}) \tag{4.1}$$

$$y'_{i,j,k} + \hat{v}_{y'_{i,j,k}} = f_{y,i,j,k}(\hat{\boldsymbol{X}}) \tag{4.2}$$

with $f_{x,i,j,k}$ and $f_{y,i,j,k}$ being the collinearity equations (cf. Subsection 2.2.3) given for reference point i, for camera j and for time point k. Observations are the pixel coordinates $x'_{i,j,k}$, $y'_{i,j,k}$

of all non-masked matched reference points from stereo image pairs as well as from independent images. Depending on the parameter setting (cf. Subsection 7.2.1), also the corresponding object coordinates are modeled as observations (cf. Subsection 2.2.3). The vector of object coordinates $\boldsymbol{X_P}$ (see also Section 2.2) is defined as

$$\boldsymbol{X_P} = \left(\boldsymbol{X}_{a,1}...\boldsymbol{X}_{a,l_a}...\boldsymbol{X}_{c,1,1}...\boldsymbol{X}_{c,k,l_{c_k}}...\boldsymbol{X}_{c,o,l_{c_o}}\right) \tag{4.3}$$

with $\boldsymbol{X}_{a,l_a}$ being the 3d object coordinates of point l_a on test field TF_a and with $\boldsymbol{X}_{c,k,l_{c_k}}$ being the object coordinates of point l_{c_k} on test field $TF_{c,k}$ for time point k of in total o time points. The interior orientation parameters are assumed to be constant over the time of image acquisition. Relative and exterior orientation parameters depend on the stereo constraint, as explained in the following.

Stereo constraint 0: Without relative orientation

No relative orientation is estimated, i.e. the two cameras are treated independently, i.e. $\boldsymbol{X}_R = \{\}$. The vector of exterior orientation parameters $\boldsymbol{X}_E$ comprises the exterior orientation parameters $\boldsymbol{X}_{E,r,k}$ and $\boldsymbol{X}_{E,s,k}$ for both the reference camera (r) and the slave camera (s) (details see *Stereo constraint 1*) for all images $1...n$ and $1...m$, respectively:

$$\boldsymbol{X_E} = \left(\boldsymbol{X}_{E,r,1}...\boldsymbol{X}_{E,r,n}\boldsymbol{X}_{E,s,1}...\boldsymbol{X}_{E,s,m}\right) \tag{4.4}$$

This stereo constraint represents the idea of not introducing any stereo rig constraints. Note the superscript o indicating that the exterior orientation parameters are given in the object coordinate system, the MCS, is omitted for visibility reasons. Either the left or the right stereo camera can be assigned as reference camera.

Stereo constraint 1: Step-wise estimation of reference and slave camera orientation

$\boldsymbol{X}_{E,r,k}$ and $\boldsymbol{X}_R$ are defined as unknown parameters for bundle adjustment. $\boldsymbol{X}_{E,s,k}$ is replaced in the collinearity equations by a formula depending on $\boldsymbol{X}_{E,r,k}$ and $\boldsymbol{X}_R$ so that observations from images of the slave camera do contribute to calibration, but no exterior orientation parameters are estimated. Thereby, this stereo constraint represents the idea of avoiding contradictions in the relative orientation by estimating the exterior orientations for only one camera, the reference camera. As disadvantage, the estimated imaging geometry of the slave camera could deteriorate, as contradictions may be become visible in the observations of this camera. The mentioned formulae to calculate the exterior orientation parameters of the slave camera are defined as

$$^{o}\boldsymbol{X}_{E,s,k} = {}^{o}\hat{\boldsymbol{X}}_{E,r,k} + \hat{\boldsymbol{R}}^{o}_{r,k} \cdot {}^{r}\hat{\boldsymbol{X}}^{s}_{r} \tag{4.5}$$

$$\boldsymbol{R}^{s}_{o,k} = \hat{\boldsymbol{R}}^{s}_{r} \cdot \hat{\boldsymbol{R}}^{r}_{o,k} \tag{4.6}$$

with ${}^{r}\boldsymbol{X}^{s}_{r}$ being the relative position between the reference camera and the slave camera given in the reference camera coordinate system and with $\boldsymbol{R}^{s}_{r}$ describing the rotation from the camera coordinate system of the reference camera r to the camera coordinate system of the slave camera s. So, $\boldsymbol{X}_R$ is defined as

$$\boldsymbol{X}_R = \begin{pmatrix} {}^{r}X^{s}_{r} & {}^{r}Y^{s}_{r} & {}^{r}Z^{s}_{r} & \theta^{s}_{r,X} & \theta^{s}_{r,Y} & \theta^{s}_{r,Z} \end{pmatrix}^{T} \tag{4.7}$$

with $\theta^{s}_{r,X}$ being the x-component of the axis angle representation of the rotation from r to s, which can be converted to $\boldsymbol{R}^{s}_{r}$. ${}^{r}X^{s}_{r}$ is the x-component of the relative position, the y- and z-components are defined accordingly. $\boldsymbol{X}_R$ is assumed to be constant over the time the image sequences are acquired.

Stereo constraint 2: Simultaneous estimation of reference and slave camera orientation

Both $\boldsymbol{X}_{E,r,k}$ and $\boldsymbol{X}_{E,s,k}$ are defined as unknown parameters for bundle adjustment. $\boldsymbol{X}_R$ is defined as in *Stereo constraint* 1, and is estimated by bundle adjustment as well. Though, in contrast to constraint 1, the relative orientation constraints are modeled as separate regular functional equations. The idea behind these additional equations is to ensure equality between the estimated relative orientation parameters and the relative orientation parameters that are calculated from the estimated exterior orientation parameters. Hereby, this stereo constraint represents the idea that contradictions may be distributed better among all parameters than in *Stereo constraint* 1. The separate functional equations are defined as

$$\mathbf{0} + \hat{\boldsymbol{v}}_{\overline{\boldsymbol{q}}_r^s} = \overline{\boldsymbol{q}}_r^s - \hat{\boldsymbol{q}}_r^s \tag{4.8}$$

$$\mathbf{0} + \hat{\boldsymbol{v}}_{r\overline{\boldsymbol{X}}_r^s} = {}^r\overline{\boldsymbol{X}}_r^s - {}^r\hat{\boldsymbol{X}}_r^s \tag{4.9}$$

with ${}^r\overline{\boldsymbol{X}}_r^s$ being the calculated relative position obtained from the estimated exterior orientations of r and s by

$${}^r\overline{\boldsymbol{X}}_r^s = \hat{\boldsymbol{R}}_{o,k}^r \cdot {}^o\hat{\boldsymbol{X}}_{E,s,k} - \hat{\boldsymbol{R}}_{o,k}^r \cdot {}^o\hat{\boldsymbol{X}}_{E,r,k} \tag{4.10}$$

and with $\overline{\boldsymbol{R}}_r^s$ describing the calculated relative orientation obtained from the estimated exterior orientations by

$$\overline{\boldsymbol{R}}_r^s = \hat{\boldsymbol{R}}_{o,k}^s \cdot \hat{\boldsymbol{R}}_{r,k}^o = \hat{\boldsymbol{R}}_{o,k}^s \cdot (\hat{\boldsymbol{R}}_{o,k}^r)^{-1} \tag{4.11}$$

To be used in the functional equations, the rotation matrix is converted to quaternions $\overline{\boldsymbol{q}}_r^s$. Quaternions have been selected for the equations for their advantageous property of not being periodic, as for example Euler angles are. So with an Euler angle of 0°, the functional equations might have zero residuals, but with the identical angle of 360°, there might be large residuals. Note again, as the relative orientation is given in the camera coordinate system of r, it does not change for different time points.

4.2.7 Datum definition for bundle adjustment for stereo cameras

Datum definition bases on the approach of Polic et al. [2018] using free adjustment, wherefore the $\boldsymbol{H}$ matrix to solve the singularity of the normal equation matrix $\boldsymbol{N}$ is required (see Subsection 2.2.4). For the proposed method, the $\boldsymbol{H}$ matrix defined by Polic et al. [2018] for a single camera is extended by additional elements for the relative orientation to support stereo camera systems.

Recall that the $\boldsymbol{H}$ matrix is derived from a similarity transform relating the inner geometry of the network to a higher-level coordinate system (see Subsection 2.2.4). Thus, $\boldsymbol{H}$ is extended as follows: As the relative rotation depends on the camera coordinate systems of r and s only, it is independent from the similarity transformation, what is reflected by the equation $\Delta\hat{\boldsymbol{q}}_r^s = \hat{\boldsymbol{q}}_r^s - {}^t\hat{\boldsymbol{q}}_r^s(\boldsymbol{q}_S) = \mathbf{0}$ with ${}^t(...)$ denoting the similarity-transformed values and $\boldsymbol{q}_S$ denoting the seven similarity transform parameters. Therefore, the corresponding partial derivatives of $\Delta\hat{\boldsymbol{q}}_r^s$ with regard to the three types of datum parameters needed for $\boldsymbol{H}$, namely translation $\boldsymbol{T}$, rotation $\boldsymbol{S}$ and scale μ, are zero as well:

$$\frac{\partial\Delta\hat{\boldsymbol{q}}_r^s}{\partial\boldsymbol{T}} = \mathbf{0} \qquad \frac{\partial\Delta\hat{\boldsymbol{q}}_r^s}{\partial\boldsymbol{S}} = \mathbf{0} \qquad \frac{\partial\Delta\hat{\boldsymbol{q}}_r^s}{\partial\mu} = \mathbf{0} \tag{4.12}$$

The difference of the relative position $\Delta^r\hat{\boldsymbol{X}}_r^s$ is assumed to be

$$\Delta^r\hat{\boldsymbol{X}}_r^s = {}^r\hat{\boldsymbol{X}}_r^s - {}^t({}^r\hat{\boldsymbol{X}}_r^s) = {}^r\hat{\boldsymbol{X}}_r^s - \mu \cdot {}^r\hat{\boldsymbol{X}}_r^s \tag{4.13}$$

This equation considers that μ is the only parameter from similarity transformation that can influence the relative position. $\boldsymbol{T}$ and $\boldsymbol{S}$ are assumed to have no influence, as the relative position is given in the camera coordinate system of the reference camera (cf. ${}^{r}(...)$), but not in the higher-level object coordinate system. This means that a translation or rotation of the imaging network (3d points, 3d camera positions and rotations) in the object coordinate system during free adjustment does not affect the relative position between the two cameras. If only μ is considered, the partial derivatives of $\Delta^r \hat{\boldsymbol{X}}_r^s$ are as follows:

$$\frac{\partial \Delta^r \hat{\boldsymbol{X}}_r^s}{\partial \boldsymbol{T}} = \boldsymbol{0} \qquad \frac{\partial \Delta^r \hat{\boldsymbol{X}}_r^s}{\partial \boldsymbol{S}} = \boldsymbol{0} \qquad \frac{\partial \Delta^r \hat{\boldsymbol{X}}_r^s}{\partial \mu} = -\boldsymbol{I}_{3x3} \tag{4.14}$$

5 Camera calibration with traffic signs

In this chapter, a method for camera self-calibration with reference points obtained from traffic signs is described. As first workflow step (Figure 5.1), semantic segmentation, coarse boundary detection and depth estimation are done for each recorded RGB image by deep learning (Section 5.1). Coarse boundaries of the traffic signs that should be used as reference information are obtained by coarse boundary detection. Second, traffic sign masks and auxiliary semantic boundaries are extracted from the semantic images (Section 5.2). Third, fine boundaries of the

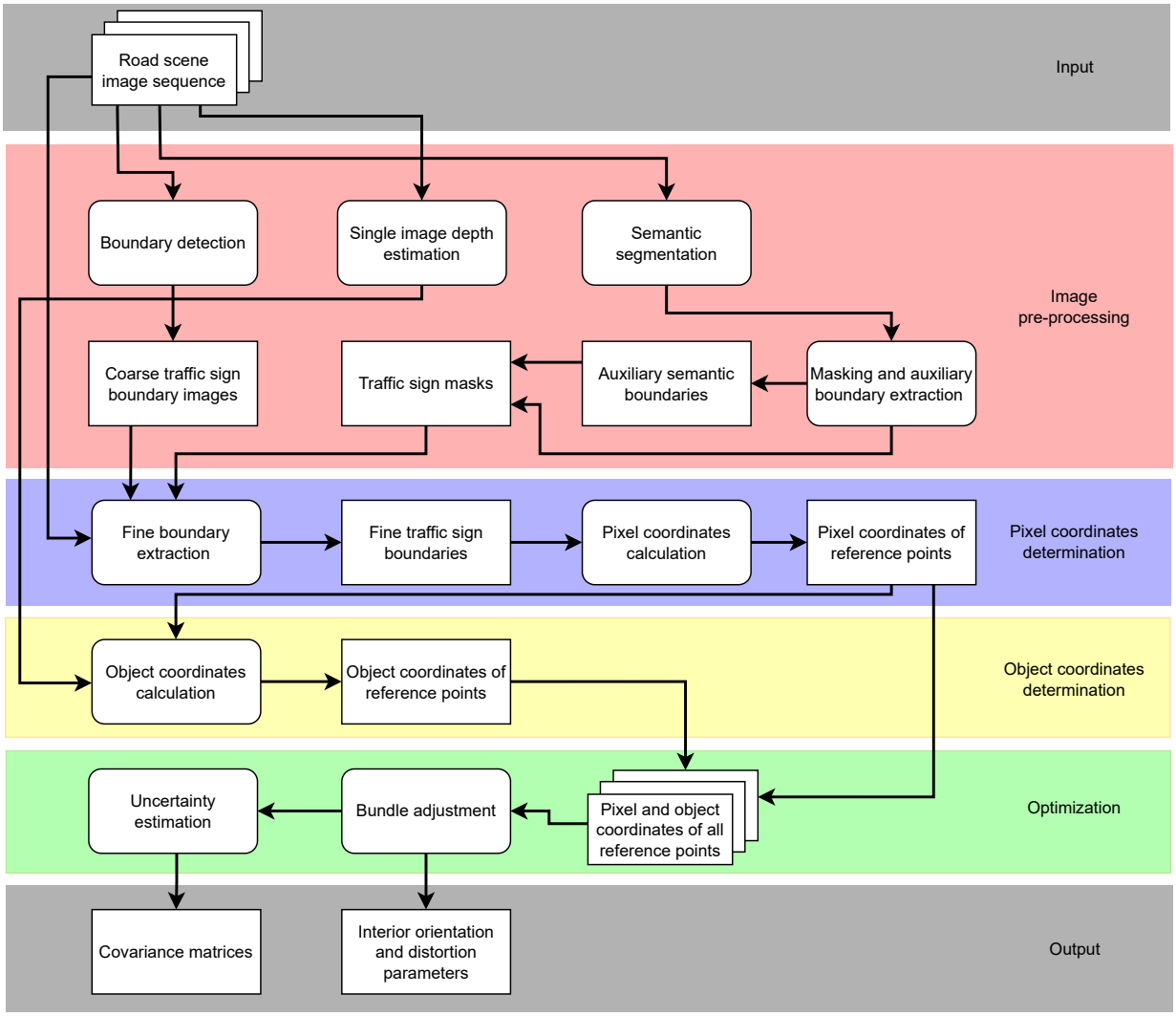
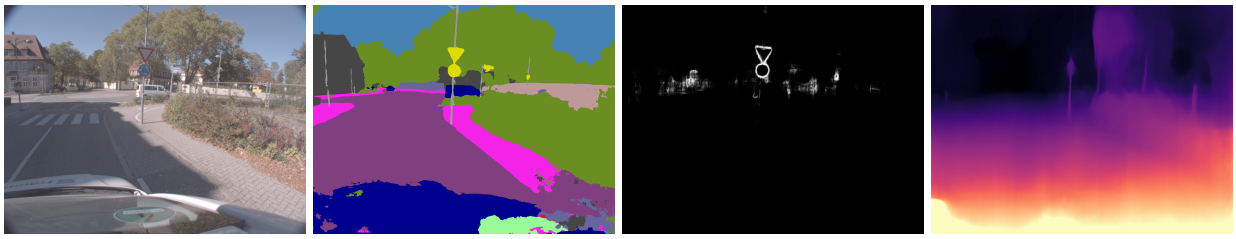


Figure 5.1: Workflow for camera calibration with traffic signs.

traffic signs are extracted using the coarse and the auxiliary semantic boundaries, as well as the RGB images. Thereby, the pixel coordinates of the reference points, which are defined by certain remarkable points along the traffic sign boundary, are calculated (Section 5.3). Forth, using the depth image and the pixel coordinates, the object coordinates of the reference points are calculated (Section 5.4). Fifth, for all images that should be used for calibration and for all processed traffic signs, the pixel and object coordinates of the reference points are processed by bundle adjustment. Thereby, the desired estimates of the interior orientation and distortion parameters and the corresponding covariance matrices are obtained (Section 5.5). This method is intended to be performed with an RGB image sequence recorded from a road scene with the vehicle camera that should be calibrated. The length of the image sequence can be determined by certain objectives, like parameter uncertainty.

5.1 Semantic segmentation, boundary detection and depth estimation

By means of deep learning, image-based semantic segmentation, boundary detection and depth estimation are performed for each acquired RGB image (Figure 5.2a). For semantic segmentation, the same method as for camera calibration by semantic structure-from-motion is used (details see Section 6.1). As most important property, the deep model needs to contain a semantic class for traffic signs, thus a pre-trained model for road scene images is used. Output is a semantic image corresponding to each RGB image (Figure 5.2b). For boundary detection, also a pre-trained model suitable for road scene images is used. It produces separate grayscale images containing the boundaries for one semantic class, whereof for the further workflow steps only the boundary images for the class *traffic sign* are used (Figure 5.2c). As these boundaries follow the traffic sign boundaries in the corresponding RGB image only roughly, they are further referred to as coarse boundaries. For depth estimation, a method that is capable to provide metric depth values, not only disparity values, is used. The depth values are provided by a so-called depth map, i.e. an image where each pixel represents a depth value, i.e. the metric distance from camera (Figure 5.2d). Advantageous, such depth maps are linked to the corresponding RGB image and so for each pixel in the RGB image a depth value can be obtained easily. Additionally, by relying on a deep model, depth estimation can be done with an image sequence from a mono camera independent from other data sources. It does neither require stereo cameras nor other vehicle sensors or receivers (e.g. GPS, IMU). It does also not require a geometric method like structure-from-motion capable for image sequences of mono cameras, which might be problematic for forward-looking cameras as the main movement is along the optical axis and so the image



(a) (b) (c) (d)

Figure 5.2: A RGB image acquired in a road scene and three images derived from it by deep learning. a) RGB image, b) semantic image showing different classes of road scene objects, among them traffic signs (yellow), c) coarse boundary image for class traffic sign, d) depth map.

geometry might be bad especially in depth direction [Vedaldi et al., 2007]. For depth estimation, also a pre-trained deep model for road scene images is used.

5.2 Masking and auxiliary semantic boundary extraction

Auxiliary semantic boundaries are extracted by contour detection [Suzuki & Abe, 1985] from the semantic images for each segment belonging to the class *traffic sign*. Hereby, each auxiliary semantic boundary is represented by a set of image points. Additionally, each segment belonging to the class *traffic sign* is binarized to obtain a binary traffic sign mask. The mask contains value 1 for all pixels belonging to class *traffic sign* in the semantic image and value 0 otherwise. As there might be multiple traffic signs visible in one RGB image, the mask may cover multiple segments. Therefore, all but the largest traffic sign segment are removed from the traffic sign mask based on the area enclosed by each auxiliary semantic boundary. This step relies on the idea, that the largest traffic sign is best suitable to determine reference points and that small segments are more likely to cause inaccurate reference point coordinates or are false positive traffic sign segments and so should be omitted. With this traffic sign mask, the boundaries of all but the largest traffic sign are removed from the coarse boundary image.

5.3 Fine boundary extraction and pixel coordinates calculation

As already mentioned, certain remarkable points along the traffic sign boundary should be used as reference points for camera calibration. For triangular and rectangular traffic signs, these are the three and four corner points, respectively. For circular traffic signs, shown in images as ellipses [Elder, 2017], these are the two end points of the major axis. To determine the pixel coordinates of these points precisely, fine boundaries of the traffic signs have to be extracted. As the semantic segments do not follow the boundaries of traffic signs in the RGB image exactly, as one segment may contain more than one traffic sign, and as also state of the art deep learning-based boundary detection provides - to the knowledge of the author - only coarse boundaries (Figure 5.2c), classic image processing is used to get fine boundaries. Furthermore, assuming that the same types of shapes are used across different countries, a classic approach may be more generic and does not require country-specific training data like a deep learning-based approach might do.

The process for fine boundary extraction is similar for the three supported shape types. In addition to the shape, only traffic signs of certain colors are supported: Circular signs and rectangular signs with blue background color (e.g. direction signs) and triangular signs with a red border (e.g. yield sign). The orientation of the shape does not play a role (e.g. upwards- or downwards-oriented triangle). For triangular traffic signs, the following major steps are employed: First, lines are extracted from the coarse boundary image by Hough transformation, which at this point contains the largest traffic sign only. Hereby it is assumed that among the extracted lines also the lines belonging to the actual boundaries of the traffic sign in the RGB image are contained. Additionally, while the coarse boundaries might be curvy, Hough transform ensures that straight lines are extracted, which corresponds better with actual traffic sign shapes. Second, all lines are associated in every possible combination to triangle candidates based on their orientation so that the sum of internal angles equals 180 degree. Third, the overlap between each triangle candidate and the traffic sign mask is determined. Candidates whose overlap is below a user-defined threshold are rejected. Candidates with an area smaller than an also user-defined minimal threshold are rejected as well. After thresholding, it is assumed that only triangle candidates remain that describe the actual triangle well with small differences in the position of the corner points. Forth, building on the assumption described in the previous sentence, the remaining triangle candidates are averaged to an intermediate triangle by K-means clustering determining three clusters, whereby each

cluster defines one corner of the intermediate triangle. Fifth, the image part corresponding to this intermediate triangle is cropped from the RGB image. The boundary of this image part is dilated by several pixels to ensure that the actual traffic sign in the RGB image is covered completely by the cropped image. Sixth, based on lower and upper color thresholds applied in the HSV color space to the cropped image, the remarkable red border of the triangular traffic sign is extracted. Last, the steps one to five are repeated with the cropped and thresholded image containing the red border only, for which it can be assumed that the outer edge of the red border matches the boundary of the traffic sign precisely. As the red color is in typical road scene images distinctive with regard to the local background on and near the traffic sign, it is considered as suitable for extracting fine boundaries. The resulting corner coordinates are the final pixel coordinates of the three reference points. For rectangles, the Hough lines are associated to rectangles, four instead of three corner points are extracted and blue instead of red color is taken. For circular traffic signs, RANSAC [Fischler & Bolles, 1981] is used to fit an ellipse to the coarse boundary image. In addition to the area threshold, ellipse candidates with large eccentricity are discarded. As according to Elder [2017], the diameter of a real-world circle corresponds with the major axis of an ellipse in an image, the end points of the major axis are selected as reference points. As with the described thresholds, bad-fitting boundaries are rejected, a preliminary shape detector determining whether a traffic sign segment belongs to a triangle, rectangle or circle sign is not necessary. At this stage of the workflow, there is a set of pixel coordinates of reference points for all successfully processed images available.

5.4 Object coordinates calculation

First, an initial guess for the focal length is obtained, then the object coordinates for the reference points are calculated. In the following, it is assumed that traffic signs in the object space are parallel to the image plane. Due to the typically large distance between traffic signs and the vehicle camera as well as the alignment of the signs towards the driver, this assumption is likely to be approximately true. Additionally, it is assumed that the principal point is equal to the image center. For the following bundle adjustment, a local object coordinate system is defined that is equal to the camera coordinate system. It means that in this coordinate system, the camera has the same exterior orientation for the entire image sequence. This definition can be interpreted as calibration with a static camera and a moving test field, though in reality the traffic signs are static and the camera-carrying vehicle is moving during image acquisition. Same as in the previous subsection, the following descriptions address triangular traffic signs, but are analogue for rectangle and circle signs.

The mentioned assumption about the principal point leads to the initial guess for the principal point coordinates, which is $(0, 0)$ in image coordinates and $(h/2, w/2)$ in pixel coordinates with h being the image height and w being the image width in pixels. The initial guess for the focal length f (cf. Figure 5.3) is obtained using an intercept theorem: It relates the known metric size of a traffic sign edge (A in the figure) and the known size of the same edge in the image (a) with the known depth (D) of a traffic sign corner point (C) to calculate an unknown auxiliary term (d). Furthermore, the angle at the projection center (O) between the lines of f and d is obtained from the sine function with the Euclidean distance between the pixel coordinates of C and the principal point as opposite leg and d as hypotenuse. Then, f can be calculated either from the inverse cosine using d and the angle or from the inverse tangent using a and the angle. For these calculations, the metric size of the edge is obtained from official regulations (cf. e.g. Department of Transport - Ireland [2010]), the size of the edge in the image is equal to the Euclidean distance between the pixel coordinates of the two corner points of this edge (Section 5.3) and the depth is obtained from the depth map (Section 5.1). To keep errors in the initial guess for the focal length resulting

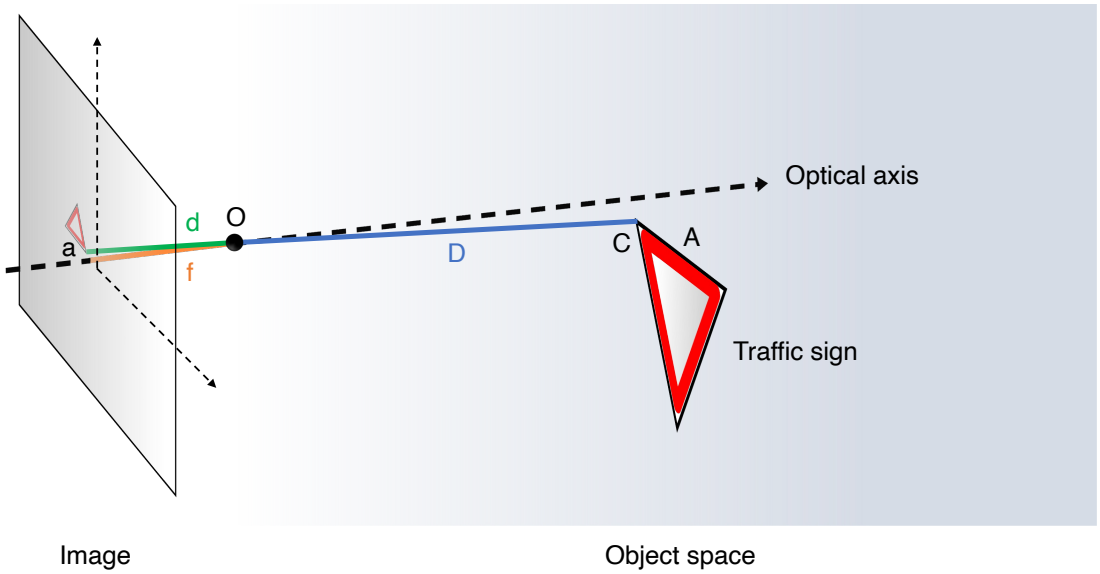


Figure 5.3: Intercept theorem used to get an initial guess for the focal length f: The distances d and D along the image ray defined by a traffic sign corner point (C) and the projection center (O), the pixel size (a) and the metric size (A) of a traffic sign edge are set in relation with f. Note that the German yield sign has rounded corners, but the reference points are defined by the imaginary corner points at the intersection of two adjacent boundaries.

from deviations of the initial guess for the principal point from the actual principal point location low, a corner point close to the optical axis should be chosen. Finally, with the pixel coordinates of C, the initial guess for the focal length and the principal point coordinates, the image ray for this reference point can be defined. With the depth value, a depth plane parallel to the image plane can be defined. Then, the object coordinates of C, defined by the intersection of the image ray with the depth plane, can be calculated. This calculation is repeated for all reference points on all traffic signs and for all acquired images, resulting in a set of object coordinates matching the previously calculated pixel coordinates.

5.5 Optimization

The next step is a global bundle adjustment, whereby the final interior and distortion parameters are estimated using the pixel and object coordinates of all reference points as observations. As the object coordinates of the reference points have metric scale, the estimated orientation parameters can be provided in metric units as well. For optimization, independent and identically distributed observation weights are assumed [Luhmann et al., 2006] as there is no a priori information available about the standard deviation of the pixel and object coordinates. The final step after global bundle adjustment is uncertainty estimation following the approach from Polic et al. [2018] (cf. Chapter 4). Thereby, the covariance matrix for the interior orientation and distortion parameters is obtained. Note again that due to the lack of external reference, the estimated uncertainty measures specify the precision instead of the accuracy.

6 Camera calibration by semantic structure-from-motion

In this chapter, a method for camera self-calibration using structure-from-motion (SfM) is presented that exploits road scene knowledge from semantic segmentation (Figure 6.1). As first step of the workflow (Figure 6.2), semantic images are obtained by semantic segmentation for all RGB images recorded in a road scene (Section 6.1). Second, for each semantic image an exclusion mask is created that consists of segments belonging to critical semantic classes, which are undesired as reference points for calibration and so the respective image parts should be excluded from feature extraction for SfM. Additional fix-pixel masks covering undesired areas visible in the same part in all images (esp. the ego-car bonnet) contribute to the exclusion masks as well (Section 6.2). Third, a 3d point cloud of the road scene is obtained by indirect SfM using the recorded road

Figure 6.1: Examples of RGB and semantic images from two points in time (t_i, t_{i+j}) from a road scene image sequence. a) RGB images, b) semantic images obtained by semantic segmentation represent different classes of road scene objects, like vehicles (blue), road (purple) or vegetation (green).

scene images (Section 6.3). Besides for the exclusion masks, the semantic images are also used to assign a semantic class to each extracted image feature in order to restrict matching to features of the same semantic class. As the object points of the point cloud are derived from matched image features and should serve as reference points for calibration, the purpose of excluding features on critical objects and restricting the matching process is to avoid undesired quality loss in calibration. After having obtained the point cloud for the entire image sequence by incremental 3d reconstruction, the Euclidean 3d point cloud of the road scene is transformed by spatial similarity transformation using filtered vehicle position data from GPS in order to incorporate a metric scale into the point cloud (Section 6.4). Forth, a global bundle adjustment is performed to obtain the final estimates for the image and object points as well as the interior and exterior camera orientation parameters (Section 6.5). By uncertainty estimation, the covariance matrices for the estimated parameters are obtained. The steps in the workflow are intended to be performed with an RGB image sequence recorded from a road scene with the vehicle camera that should be calibrated. The image sequence can cover a time span as desired by an application or a certain objective, like parameter uncertainty, and recording can be repeated also as desired.

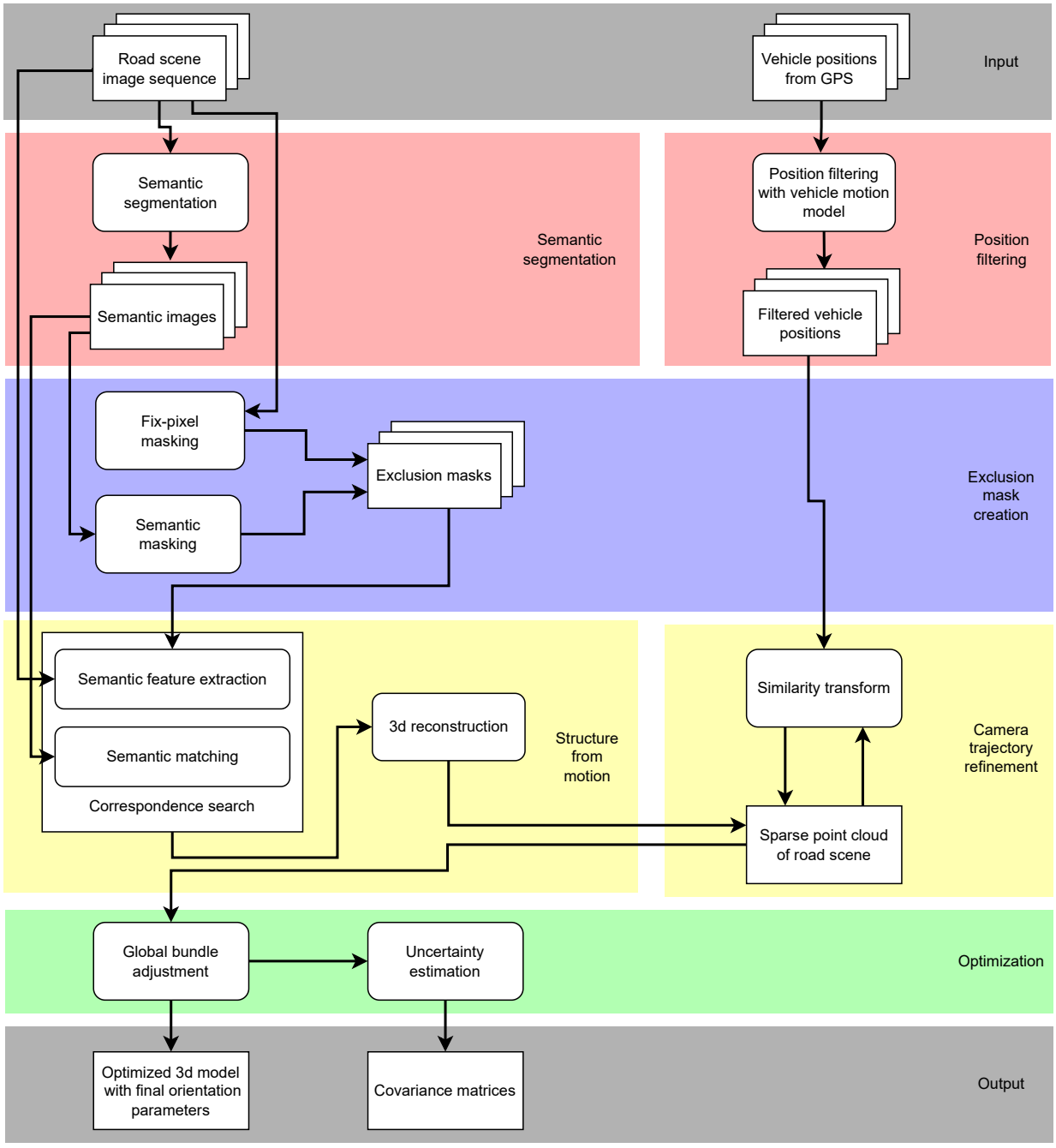


Figure 6.2: Workflow of the proposed method for self-calibration of a vehicle camera by structure-from-motion using scene knowledge obtained from semantic segmentation.

6.1 Semantic segmentation

Image-based semantic segmentation is applied to obtain semantic knowledge about the road scene. Instance segmentation is not used in the proposed method, as image features may occur in all image parts and instance segmentation often covers only certain classes shown in some image parts (see Section 2.3). For segmentation, inference with a trained deep model is done for each RGB image (Figure 6.1a) of the calibration image sequence, resulting in pixel-wise semantic images (Figure 6.1b). A model is used that has been trained on road scene images with ground truth annotations distinguishing common semantic classes for road scenes like *road*, *vehicle* or

building. By using a model trained by a third party [e.g. Chen et al., 2018b], the need for own time-intensive hyperparameter tuning and expensive high performance computing capabilities required for large-scale training datasets is avoided (e.g. 370 GPUs as reported by Chen et al. [2018a]). Furthermore, by using a model trained on appropriate third-party datasets, the often costly need to acquire ground truth annotations for own image sequences can be avoided and the risk of overfitting the model to own datasets is eliminated.

6.2 Exclusion mask creation

Semantic masking is applied during SfM in order to exclude image parts showing potentially critical objects from feature extraction to avoid unreliable reference points for camera calibration. Hence, features can be extracted only from image parts that are considered as non-critical. Hereby, the approach is to identify potentially critical objects by their semantic class. The binary exclusion masks used to implement semantic masking are derived from the semantic images, whereby one intensity value is assigned to pixels belonging to semantic classes which are considered non-critical and should be used for feature extraction (white color in Figure 6.3a). The other intensity value is assigned to pixels belonging to semantic classes that are considered as critical and should not be used for feature extraction (black color in Figure 6.3a). The list of critical semantic classes has to be provided manually before calibration and can cover for example all moving objects or objects with a reflecting surface; a comparison between different classes with regard to the resulting distribution and frequency of reference points can be found in Hanel & Stilla [2019b].

(a) (b)

Figure 6.3: Examples (corresponding to Figure 6.1) for exclusion masks and for semantic matching. a) Exclusion masks for classes vehicle and ego-car. The masked area (black color) on the ego-car is obtained by fix-pixel masking, while the masked area on the two other vehicles is obtained by semantic masking, b) semantic matches that are only allowed between feature points of the same semantic class, e.g. traffic signs (yellow) or sidewalk (pink).

In addition to semantic masking, critical objects shown in the same part of the image in the entire image sequence, like the ego-car, are excluded by fix-pixel masking. Therefore, a mask consisting of polygons defined by fix image points is manually determined before calibration and added to the exclusion masks (Figure 6.3a). The ego-car is considered as critical object, as the windshield or the bonnet can easily show reflections and lead to wrong feature correspondences.

6.3 Structure-from-motion

Semantic feature extraction

SIFT features [Lowe, 1999] are extracted from each image of the road scene image sequence using the exclusion masks in order to obtain feature points only on such road scene objects that are not considered as critical.

Semantic feature matching

The idea of semantic feature matching is to perform matching only between features extracted from image parts showing objects from the same semantic class in order to reduce obviously wrong matches between objects of different types, like between vehicle and building. Before matching, the semantic class is assigned to each extracted feature point by sampling at the feature location in the corresponding semantic image. During matching, the semantic classes of both features being the current match candidates are compared with each other and regular matching is done only if they belong to the same semantic class (Figure 6.3b). Hereby, regular matching means that a descriptor similarity measure decides whether the two candidates are accepted as match or not. If the candidates are from different classes, processing continues with the next match candidates. Semantic masking, fix-pixel masking and semantic matching can be applied during the workflow either separately, in combination or even not.

3d reconstruction

Having feature correspondences established by feature extraction and matching, a sparse 3d point cloud of the road scene is obtained iteratively by 3d reconstruction. The reconstruction is initialized with a random image pair, then step-by-step the other images are registered, 3d coordinates of object points calculated by triangulation and optimized by bundle adjustment. For the reconstruction, the same set of interior and distortion parameters is used for all images, assuming that there have been no changes in these parameters while recording the image sequence. This assumption has the beneficial effect that the point cloud is Euclidean already [Hartley, 1993] and only lacking scale information to become metric Heyden & Astrom [1996].

6.4 Position filtering and camera trajectory refinement

The metric scale information is incorporated into the 3d point cloud using a similarity transformation based on filtered and refined 3d camera positions, wherefore the fact is used that the positions of the mono camera (obtained from the exterior orientations) in the point cloud can be uniquely assigned to the vehicle positions obtained from GPS. Even though the datum point of the vehicle positions is typically different from the camera position (e.g. the center of the front axis), the distance between two camera positions and thus the metric scale is not influenced and so the differences in the datum point definition can be neglected.

The initially calculated metric GPS positions of the vehicle are filtered by special vehicle motion models [Schubert et al., 2008; Hanel et al., 2019], e.g. based on the Ackermann movement model or on non-holonomic constraints [Scaramuzza et al., 2009; Ruland et al., 2010; Lee et al., 2013]. By exploiting the fact that vehicle motion is restricted to certain degrees of freedom (e.g. planar movement) it is intended to mitigate the effect of observation errors in the GPS positions in order to get a smoother trajectory that is appropriate for vehicle motion patterns and to increase the quality of the metric scale and so the quality of camera calibration subsequently. In the first step of trajectory refinement, the initially calculated GPS vehicle positions are interpolated to match the image acquisition time points. As the positions from GPS are typically available with a remarkably higher frequency than the images, a linear interpolation is considered to be sufficient. In the second step, Kalman filtering based on the vehicle motion model is applied to the interpolated trajectory in order to get a filtered trajectory (details see Hanel et al. [2019]). For numerical benefits in the subsequent optimization step, the center of the filtered trajectory is placed in the origin of the object coordinate system of the 3d reconstruction. Finally, the transformation parameters are calculated based on the filtered and refined vehicle trajectory. By applying the transformation to the entire point cloud, the metric scale gets incorporated into it.

6.5 Optimization

In the last step, the final interior orientation parameters are estimated by global bundle adjustment covering all images of the sequence. The corresponding covariance matrices are obtained as uncertainty measures afterwards. For details, see Section 5.5.

7 Test data and experiments

This chapter covers descriptions of the datasets used and the experiments performed.

7.1 Datasets

The *Stereo image dataset* (Subsection 7.1.1) is used for stereo camera calibration with test fields (Chapter 4). The *Ettlingen sequence* and the *Munich sequence* (Subsection 7.1.2) are used to test the methods for self-calibration with traffic signs (Chapter 5) and for self-calibration by semantic structure-from-motion (Chapter 6). The *Fraunhofer calibration dataset* (Subsection 7.1.3) is used to provide a reference calibration to evaluate the self-calibration methods.

7.1.1 Stereo image dataset

The *Stereo image dataset* contains pairs of simultaneously taken stereo images showing two 2d test fields, one with coded and uncoded reference marks from the manufacturer Aicon [Schneider et al., 2017], the other with a classic checkerboard pattern (Figure 4.3). The points in the center of the reference marks and the corner points of the checkerboard pattern define the reference points for calibration. The dataset has been recorded by a pair of monochromatic industrial cameras (Figure 4.1, specifications see Table 7.1). The baseline between the cameras is approximately 33 cm in length and is oriented approximately orthogonal to the optical axes of the cameras. The relative orientation and base length are typical for automotive camera systems (e.g. 22 cm [Cordts et al., 2016], 30 cm, 53 cm, 57 cm [Rehder et al., 2017]). Likewise, the chosen focal length of 6 mm is typical for automotive cameras, e.g. Rehder et al. [2017] are using optics with focal length values of 4 mm, 4.5 mm and 15 mm for their experiments. The dataset comprises two image sequences, each around 30 images, one taken in a car with the windshield in front of the cameras, the other taken in a lab without a windshield (cf. Figure 4.3).

Table 7.1: Specifications of cameras and optics used for the experiments.

Stereo image dataset	Ettlingen, Munich and Fraunhofer datasets
Camera: SVS-VISTEK SVCam eco655MVGE	Baumer VLG-20C.I.
Monochrome CCD	Color CCD
2448 x 2050 px	1624 x 1228 px
3.45 x 3.45 μm	4.4 x 4.4 μm
Optics: VS Technology SV-0614H	-
Focal length 6 mm	-
Aperture 1.4 $\sim$ 16	-
SVS-Vistek GmbH [2020]	Borgmann et al. [2018]
VS Technology [2015]	Baumer GmbH [2019]

7.1.2 Ettlingen and Munich datasets

The *Ettlingen sequence* (Figure 7.1a) consists of 300 RGB road scene images recorded by the multi-sensor vehicle MODISSA [Borgmann et al., 2018] during a 30 second drive through a suburban area. MODISSA (Figure 7.2a) is operated by the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation and is equipped with a range of different environment-perceiving sensors, of which the on-board front-right forward-looking camera (specifications see Table 7.1) is used to record the sequence. Additionally, the trajectory of MODISSA is recorded with a GPS receiver; as the recording times of the GPS positions and images do not match, they need to be synchronized to each other.

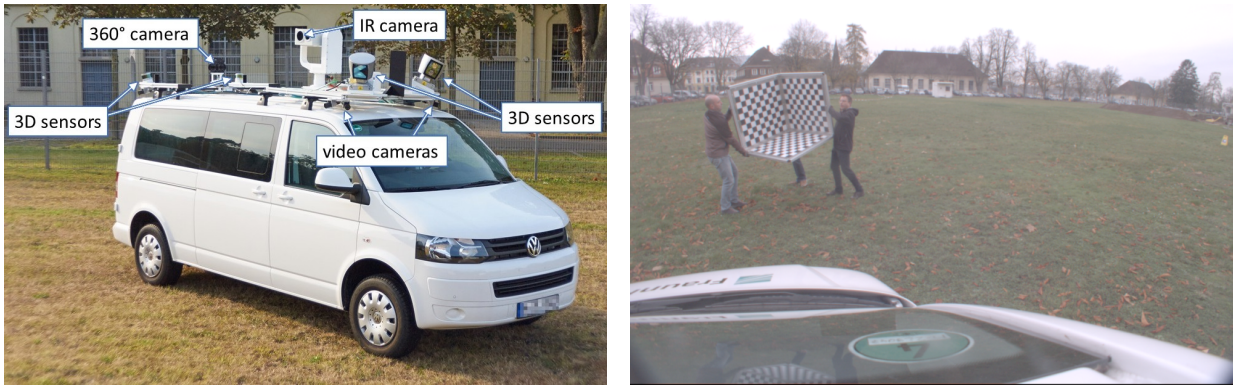
The *Munich sequence* (Figure 7.1b) consists of 504 road scene images and has been recorded also by MODISSA during a 40 second drive through an urban area. Compared to the *Ettlingen sequence*, the *Munich sequence* shows a larger number of traffic signs, especially at the two intersections contained in the sequence. Furthermore, it contains more moving objects like cars and pedestrians.

(a) (b)

Figure 7.1: Example images of the *Ettlingen sequence* and *Munich sequence*. a) Ettlingen sequence recorded in a suburban environment, b) Munich sequence recorded in an urban environment.

7.1.3 Fraunhofer calibration dataset

The *Fraunhofer calibration dataset* consists of an image sequence (example see Figure 7.2b) recorded at the Fraunhofer research facility with the same forward-looking on-board camera of MODISSA as the *Ettlingen sequence* and the *Munich sequence*. The *Fraunhofer calibration dataset* contains more than $1,000$ images showing a 3d calibration test field being moved and rotated through the field of view of the camera. The test field has been constructed by three orthogonal planes forming an "open cube" with checkerboard patterns on each plane. The checkerboard corners define the reference points for calibration. As according to Luhmann et al. [2016], calibration with 2d test fields can be disadvantageous in terms of accuracy and correlations between parameters, a 3d shape has been selected over a 2d shape for the test field. The object coordinates of the reference points have been determined analytically assuming a grid shape and a known metric checkerboard square size.



(a) (b)

Figure 7.2: Sensor setup and example image from a camera of the multi-sensor vehicle MODISSA. a) MODISSA. The front-right camera named "video camera" has been used to record the image sequences for the *Munich*, *Ettlingen* and the *Fraunhofer calibration dataset*. Image adapted from Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB [2020], b) example image from the *Fraunhofer calibration dataset* showing the 3d calibration test field that is rotated and moved in the field of view of the camera of MODISSA.

7.2 Experiments

This section covers the description of the experiments performed to test the proposed methods, including the definition of experimental cases that allow to investigate different aspects of the methods and also including technical details on the implementation. First, the experiments for the method for test field calibration are addressed (Subsection 7.2.1), followed by the experiments for self-calibration with traffic signs (Subsection 7.2.2). Last, the experiments for self-calibration by semantic structure-from-motion are explained (Subsection 7.2.3).

7.2.1 Camera calibration with test fields through a vehicle windshield

First, the experiments for stereo camera calibration with test fields are addressed.

Experimental cases and parameter settings

The calibration is performed with the *Stereo image dataset* for all possible combinations of experimental cases and parameter settings (Table 7.2). Hereby, experimental cases address major evaluation aspects to answer the research questions: The *feature types* describe the type of reference marks used, either only from the test field with Aicon reference marks (feature type *Aicon*), or only from the checkerboard test field (*Checkerboard*) or jointly from both test fields (*Merged*). The camera models are used for calibration as defined in Subsection 2.1.4 and the stereo constraints are used as defined in Subsection 4.2.6. The two setups represent calibration with and without the windshield. In addition to the experimental cases, parameter settings address minor aspects with influence on camera calibration: The idea behind keeping the object coordinates of the reference points fixed during optimization (*3d points fixed*) is that they have been predetermined with high accuracy, and allowing updates to them may undesirably cover effects from the windshield, which otherwise could become visible in the estimated orientation parameter values. Robust optimization could help to alleviate the influence of potential outliers in the observations in bundle adjustment and has been introduced as consequence of preliminary evaluations (details see Subsection 8.2.5). As the relative orientation constraints in *Stereo constraint 2* are defined as fictional observations with regular functional equations, they have observation

Table 7.2: Experimental cases and parameter settings for stereo camera calibration with test fields.

Experimental cases				Parameter settings		
Setups	Camera models	Feature types	Stereo constraints	3d point optimization	Optimization mode	Observation weights real:fictional
with	*None*	*Aicon*	*0*	*Adjustable*	*Robust*	*1:10*
without	*Radial*	*Checkerboard*	*1*	*Fixed*	*Non-robust*	⋮
	Both	*Merged*	*2*			*10:1*

weights. Therefore, the grade how strict these constraints are enforced during bundle adjustment can be modeled by different ratios of observation weights between these fictional and the real observations (pixel coordinates and, if applicable, object coordinates). The left stereo camera is defined as reference camera, the right one as slave camera (cf. Subsection 4.2.6). Finally, note that a short nomenclature based on the initial letters is introduced for the experimental settings: E.g. "RM0" means that the *Radial* camera model, feature type *Merged* and *Stereo constraint 0* are used.

Implementation and execution

The pixel coordinates of the Aicon reference marks are extracted with the software Aicon 3D Studio [Schneider et al., 2017]. The pixel coordinates of the checkerboard corners are extracted using Matlab-internal routines. 3d reconstruction and bundle adjustment are done with the framework COLMAP [Schönberger & Frahm, 2016] . Uncertainty estimation is done with the framework USfM [Polic et al., 2018]. Note that COLMAP, VisualSFM [Wu, 2013] and other common 3d reconstruction frameworks do not provide uncertainty information to the best knowledge of the author and so a separate solution becomes necessary. All interior, relative and exterior orientation parameters are set as adjustable during bundle adjustment. 3d points are set either as adjustable or as fixed, cf. parameter settings. Observations for bundle adjustment are the pixel and, also depending on the parameter settings, the object coordinates of the reference points. Optimization is terminated when the convergence criterion has been met (change in the sum of absolute residuals below a pre-defined threshold or after 10,000 iterations). COLMAP and USfM have been adopted and extended to execute the proposed algorithm. The remaining steps of the workflow are done by own Matlab or Octave routines. Calibration is performed separately for both setups. The initial guess for the interior orientation and distortion values needed for a Euclidean 3d reconstruction and non-linear optimization is obtained from a geometric camera calibration having been performed beforehand separately for each camera. Interior orientation and lens distortion are assumed to be constant over the entire image sequence, which is a common assumption in the automotive domain, as regularly lenses with fixed focal length and focus setting are used in vehicles [Gil et al., 2018b]. The initial guess for the relative orientation is obtained from a geometric stereo calibration. Calibration is repeated 25 times for each experimental case and parameter setting to alleviate potential non-deterministic effects (cf. second paragraph in Section 8.2).

7.2.2 Camera calibration with traffic signs

Second, the experiments for camera calibration with traffic signs are addressed.

Experimental cases

The experimental cases (Table 7.3) cover all combinations of supported traffic sign shapes for the most-promising semantic segmentation and depth estimation methods (cases *1* ...), which have

been selected based on preliminary experiments (see Subsection 8.3.4). They also cover other less-promising semantic segmentation and depth estimation methods, but only for the combination with all three traffic sign shapes (cases *2 CRT* and *3 CRT*).

Table 7.3: Experimental cases for camera calibration with traffic signs.

Case	Traffic sign shapes	Computer vision methods
1 C	Circle	Deeplabv3+, Monodepth2
1 R	Rectangle	Deeplabv3+, Monodepth2
1 T	Triangle	Deeplabv3+, Monodepth2
1 CR	Circle, rectangle	Deeplabv3+, Monodepth2
1 CT	Circle, triangle	Deeplabv3+, Monodepth2
1 RT	Rectangle, triangle	Deeplabv3+, Monodepth2
1 CRT	Circle, rectangle, triangle	Deeplabv3+, Monodepth2
2 CRT	Circle, rectangle, triangle	Deeplabv3+, struct2depth
3 CRT	Circle, rectangle, triangle	EfficientPS, Monodepth2

Implementation and execution

Camera calibration with traffic signs is tested with the *Ettlingen* and the *Munich sequence*. Semantic segmentation is performed with either Deeplabv3+ [Chen et al., 2018b] or with EfficientPS [Mohan & Valada, 2021]. Depth estimation is either performed with Monodepth2 [Godard et al., 2019] or with struct2depth [Casser et al., 2019], which have been selected as they are recently published methods for which the source code is available to the public. Deep models trained (e.g. on the well-known road scene image dataset KITTI as done by Casser et al. [2019]) and provided by the authors of these methods are used for the experiments. Fine tuning on the two test datasets has not been done to avoid overfitting. Hence, it is possible to show with the experiments that the proposed method does not only work with images used for training already, which would not be realistic for an automotive application where future images will be acquired in completely different road scenes and conditions. The *Both* camera model with radial and tangential distortions is used (cf. Subsection 2.1.4), as it has shown best results in calibration with test fields (see Subsection 9.1.1). A single set of interior orientation parameters is estimated for each entire image sequence assuming no changes to the interior orientation during the few seconds of data recording. Bundle adjustment and uncertainty estimation are done with COLMAP and USfM.

Reference calibration

The proposed methods for calibration with traffic signs and semantic structure-from-motion are evaluated against a reference calibration performed as high-quality test field calibration with the *Fraunhofer calibration dataset*. The 3d object coordinates of the reference points are defined in a coordinate system with the intersection point of the three planes as origin and with each intersection line between two of the three planes as coordinate axis. The coordinate values are calculated assuming an ideal grid shape of the reference points and using the known edge length of one checkerboard square. Bundle adjustment is done with COLMAP and uncertainty estimation with USfM as well. The relevant results of reference calibration are the estimated values and standard deviations of the interior orientation and distortion parameters.

7.2.3 Camera calibration by semantic structure-from-motion

Third, the experiments for camera calibration by semantic structure-from-motion are addressed. For comparison of the proposed method with a high-quality reference calibration, the same reference calibration as for calibration with traffic signs is utilized. Several experimental cases are defined covering different approaches to integrate semantic knowledge and a vehicle motion model.

Experimental cases

The experimental cases (Table 7.4) cover all combinations where in the workflow either no semantic knowledge, only semantic feature extraction (*SFE*), only semantic matching (*SM*) or both are used. They also cover all combinations where either no special vehicle motion model or where a special vehicle motion model (*VMM*) is applied to filter and refine the GPS positions of the vehicle. For experimental cases with VMM, only the later part of the workflow is executed by taking the 3d reconstruction from the corresponding experimental case without VMM and continuing with the step of camera trajectory refinement (see Section 6.4). For semantic feature extraction, 10 different mask types (MT) are derived from semantic categories (Table 7.5). The semantic categories consist of either a single or of multiple semantic classes following the Cityscapes class definition [Cordts et al., 2016]: (i) *vehicle: bicycle, bus, car, caravan, license plate, motorcycle, trailer, train, truck*, (ii) *nature: terrain, vegetation*, (iii) *human: person, rider*, (iv) *construction: bridge, building, fence, guard rail, tunnel, wall*, (v) *flat: parking, rail track, road, sidewalk*, (vi) *object: pole, pole group, traffic light, traffic sign* and (vii) the class *sky* forms its own category. The only category not depending on semantic road scene knowledge is *ego-car* originating from fix-pixel masking. As mask type 1 (MT1) is an empty mask, semantic feature extraction has no effect (therefore, "+ SFE" is not added to MT1 in Table 7.4) and so experimental cases with MT1 can be seen as baseline for comparisons between the mask types. Furthermore, in combination with SM, MT1 defines the experimental case where only semantic feature matching is used.

Table 7.4: Experimental cases for calibration by semantic structure-from-motion. Semantic feature extraction (SFE) is applied to all except the empty mask type 1 (MT1), which can be seen as baseline for comparison with other mask types. Semantic matching (SM) and the vehicle motion model (VMM) are applied in all four combinations.

	Without vehicle motion model	With vehicle motion model
Without semantic matching	MT1	MT1 + VMM
	MT2 + SFE	MT2 + SFE + VMM
	⋮	⋮
	MT10 + SFE	MT10 + SFE + VMM
With semantic matching	MT1 + SM	MT1 + SM + VMM
	MT2 + SFE + SM	MT2 + SFE + SM + VMM
	⋮	⋮
	MT10 + SFE + SM	MT10 + SFE + SM + VMM

Implementation and execution

Calibration by semantic structure-from-motion is tested with the *Ettlingen* and the *Munich sequence*. Semantic segmentation is performed with the Deeplabv3+ network [Chen et al., 2018b], which is, despite its age and the fast research progress in deep learning, still among the best methods in the Cityscapes benchmark for semantic segmentation of road scene images [Cordts et al., 2019]. A model for this network trained on the Cityscapes dataset [Cordts et al., 2016]

Table 7.5: Mask types (MT) for semantic feature extraction.

Mask type	Semantic categories
MT1	none (empty mask)
MT2	*ego-car* + *vehicle*
MT3	*ego-car* + *nature*
MT4	*ego-car* + *sky*
MT5	*ego-car* + *human*
MT6	*ego-car* + *construction*
MT7	*ego-car* + *flat*
MT8	*ego-car* + *object*
MT9	*ego-car* + all road users (*human, vehicle*)
M10	*ego-car* + all movable objects (*human, sky, vegetation, vehicle*)

and provided by the Deeplabv3+ developers is used; i.e, the model is not fitted specifically to the test data, same as for calibration with traffic signs. Panoptic segmentation [Mohan & Valada, 2021] providing semantic information for each individual object, whereby multiple objects may be covered in one segment in semantic segmentation, would have been an interesting alternative for integrating semantic knowledge. But as the experiments have been carried out before publication, it could not have been considered. 3d reconstruction by structure-from-motion is done with COLMAP [Schönberger & Frahm, 2016], which has shown best performance in a comparison with other structure-from-motion algorithms [Bianco et al., 2018]. The *Both* camera model is used (cf. Subsection 2.1.4). A single set of interior orientation parameters is estimated for each image sequence assuming no changes to the interior orientation during the few seconds of data recording. COLMAP's "sequential feature matching" is applied considering the fact that images recorded during driving are already in a temporarily sequential order. By this matching algorithm, only the previous and next twenty images are considered. Integration of the filtered vehicle positions as well as global bundle adjustment is also done by COLMAP. The uncertainty measures are estimated by USfM [Polic et al., 2018]. Experiments where bundle adjustment did not converge after a pre-defined number of optimization iterations are discarded and the workflow is repeated from scratch. Preliminary empiric analysis has shown unfavorable initial image pairs for 3d reconstruction obtained by random selection to cause that convergence is not achieved. According to this analysis, the undesired effect of such experiments are residuals that are more than hundred times larger than for a converging optimization with a better selection of the initial image pair.

8 Results and discussion

This chapter covers the results and discussions. First, the evaluation approaches with statistical measures, deviation plots and statistical significance tests are described (Section 8.1). Second, the evaluation of camera calibration with test fields through a vehicle windshield is addressed (Section 8.2). Due to their high relevance for the evaluation of this method, residuals of the image points are shown as plots besides numerical representation of other statistical measures. The deviation plots and significance tests compare the interior, relative and exterior orientation in the two calibration setups with and without the windshield with each other. Furthermore, correlations among and between the interior and relative orientation parameters are shown and compared between the two setups. Remarkable observations from these results are discussed. Third, the evaluation of camera calibration with traffic signs is addressed (Section 8.3). Besides the statistical measures, deviation plots and significance tests comparing the proposed method with a reference calibration are shown. Potential factors influencing the calibration are discussed. Fourth, the evaluation of camera calibration by semantic structure-from-motion is presented (Section 8.4), which is analogue to the evaluation of calibration with traffic signs. Fifth and last, the three proposed methods and their results are discussed in comparison with each other (Section 8.5).

8.1 Evaluation approaches

All proposed methods are evaluated with the following approaches: First, statistical measures like the number of extracted features (#features), the number of image points (#image points) and the number of object points (#object points) of the reference points give basic insights into the image network and reference information used for calibration. Among these measures, large values are desirable for camera calibration. Additionally, the mean values $\bar{r}_x$, $\bar{r}_y$ and standard deviations σ_{r_x}, σ_{r_y} of the residuals of image points in x- and y-direction are shown. In contrast, small values are desirable for these measures. Large mean residuals could be an indicator for systematic biases in the observations, while large standard deviations could be an indicator for low precision of the observations. The value range of the residuals should be equal in the x- and y-direction as large differences could be an indicator for direction-dependent systematic effects, for instance due to incomplete DOF coverage in self-calibration. Second, deviation plots of the estimated interior, and if applicable, relative and exterior orientation parameter values and standard deviations between the proposed method and a reference calibration are created. Objective of these plots are visual analysis of differences between orientation parameters, experimental cases and test sequences. Small deviations from reference calibration are desirable for the estimated parameter values, which can be an indicator for a reliable calibration with the proposed methods. Either negative or at least small positive deviations are desirable for the estimated standard deviations, which can be an indicator that the proposed methods either have a better or a similar precision as the reference calibration. Third, the statistical significance of the mentioned deviations is assessed by hypothesis tests. The tests are complementary to the plots: While the tests allow for conclusions on the relevance of deviations, the plots allow for conclusions on the direction and

strength of deviations. Finally, remarkable observations from the plots, tables and tests about the results are discussed. Such observations represent tendencies in the results that are valid for a majority of experimental cases, orientation parameters, etc., despite that there might be single experimental cases, parameters, etc. that show contradictory results.

One drawback of the presented evaluation is that the standard deviations of the orientation parameters represent the inner accuracy only, also called precision. Hence, they indicate how well the observations, the functional model and the estimated parameter values match with each other [Luhmann et al., 2006, 2013]. But due to the lack of additional independent reference points or reference lengths with higher accuracy than the observations, the standard deviations don't represent the outer accuracy, so they don't indicate how well the estimated parameters coincide with their actual values [Remondino et al., 2017]. As worst-case example, the estimated focal length may be twice as large as the actual value due to a scale error that affects all observations, but the estimated standard deviation of the focal length can be misleadingly small, if the functional model, observations and estimated parameters still do match well due to the equal effect of the scale error. Another drawback is that two contra-intuitive effects have to be acknowledged when interpreting the results of the hypothesis tests on the estimated parameter values. First, if the standard deviation of an estimated parameter is large, the hypothesis test on the deviation between the estimated and the reference value may decide for *non-significant* even if the deviation is so large that the estimated value will not be considered as reliable by visual inspection. Second, if the standard deviation is small, the same hypothesis test may decide that the deviation is *significant*, even if it is so small that the estimated value will be considered as visually reliable.

8.2 Camera calibration with test fields through a vehicle windshield

For calibration with test fields, the evaluation is more comprehensive than described in the previous section as the influence of the windshield should be analyzed from different perspectives. It comprises (i) statistical measures and residual plots (Subsection 8.2.1) and (ii) deviation plots between the setup with and without the windshield for the estimated interior and relative orientation parameter values. Furthermore, it comprises deviation plots for the estimated standard deviations of the interior, relative and exterior orientation parameters between the two setups (Subsection 8.2.2), (iii) statistical significance tests for the mentioned deviations (Subsection 8.2.3), (iv) plots of the correlations among and between the interior, relative and exterior orientation parameters (Subsection 8.2.4) and (v) a discussion of remarkable observations (Subsection 8.2.5).

Each experimental case was repeated in total 25 times (called *experimental runs*) to alleviate the influence of non-deterministic effects, for example during optimization [Agarwal et al., 2022]. If not stated otherwise, the data shown for each experimental case is the average over all experimental runs. As optimization with the *None* camera model was not successful for the *without* setup, no data is shown for this camera model; the lack of distortion parameters in only this camera model might be a reason for the failing optimization. For each camera, the standard deviations of the three position and three rotation parameters of the exterior orientation of all images are averaged to a single standard deviation value for the position and a single standard deviation value for the rotation, as the object coordinate systems in both setups are not identical and so a comparison based on the three individual parameters would not be meaningful. For the same reason, only the standard deviations of the exterior orientation parameters are considered for evaluation, but not the orientation parameter values.

Outliers in the estimated parameters and standard deviations are removed prior to evaluation based on the median absolute deviation (MAD). Classic outlier removal based on mean and stan-

dard deviations (e.g. 3σ rule) is considered as not robust, as both mean and standard deviations are sensitive to outliers [Leys et al., 2013]. As stringency threshold for the MAD, the value 10 is selected, i.e. the chance for considering a parameter value or standard deviation as outlier can be seen as small [Leys et al., 2013].

8.2.1 Statistics and residual plots

Statistical measures for both setups, all feature types and two camera models are shown in Table 8.1. The number of images used for calibration (#images) is the sum over both cameras. As it can be expected due to the combination of both test fields, the number of images, image and object points of the reference points is highest for the feature type *Merged* in both setups. The higher numbers for the feature type *Aicon* compared to *Checkerboard* correspond with a larger number of images and with a larger number of reference marks on this test field. The higher numbers of image and object points for the *with windshield* setup correspond with a higher number of images in this setup from which reference points have been extracted successfully.

Table 8.1: Statistical measures for both setups and all experimental cases for calibration with test fields. Measures shown for all feature types (A: *Aicon*, C: *Checkerboard*, M: *Merged*) and for two camera models (B: *Both*, R: *Radial*). No distinction between stereo constraints, as the values are the same for all stereo constraints.

Setup	Experimental case	#images	#image points	#object points
Without windshield	RM	40	62,390	282
	BM	40	62,390	312
	RA	35	52,900	146
	BA	35	52,900	146
	RC	21	9,490	54
	BC	21	9,490	54
With windshield	RM	62	144,232	443
	BM	62	144,232	419
	RA	48	113,938	147
	BA	48	113,938	146
	RC	34	30,294	54
	BC	34	30,294	58

Histogram plots of the residuals of the image points in x- and y-direction for both cameras are shown separately for each setup (Figures 8.1, 8.2). The histograms are normalized over the number of parameter settings and experimental runs, so that each histogram covers all images taken for one camera in one experimental case. The horizontal axis of the histograms is limited to the $\pm 3.5\sigma$ interval, the tick marks match 1σ, 2σ and 3σ. As coherent observations for both setups, it can be seen that the residuals are in average clearly larger for *Stereo constraint 1* than for the other two stereo constraints (larger scaling on horizontal axes). Between the other two stereo constraints, no differences are visible (same scaling on horizontal axis) in either setup. The steep curves for *Stereo constraint 1* indicate the presence of a large number of outliers. For both setups, the residuals are larger for camera model *Both* than for *Radial* (larger scaling). Additionally, mean biases and negative skew can be seen especially for the *Both* camera model in the *with windshield* setup. Skew is larger for the y-direction than for the x-direction (cf. especially the first row in Figure 8.2). Remarkable differences between the two cameras can be only seen in a few experimental cases (e.g. *without windshield*, RA1). Furthermore, residuals are larger for the *with windshield* setup in all experimental cases (larger scaling on horizontal axis). While the residuals in the *with windshield* setup are in a reasonable value range for *Stereo constraints 0* and *2* ($1\sigma \leq 1px$), this does not apply for *Stereo constraint 1*.

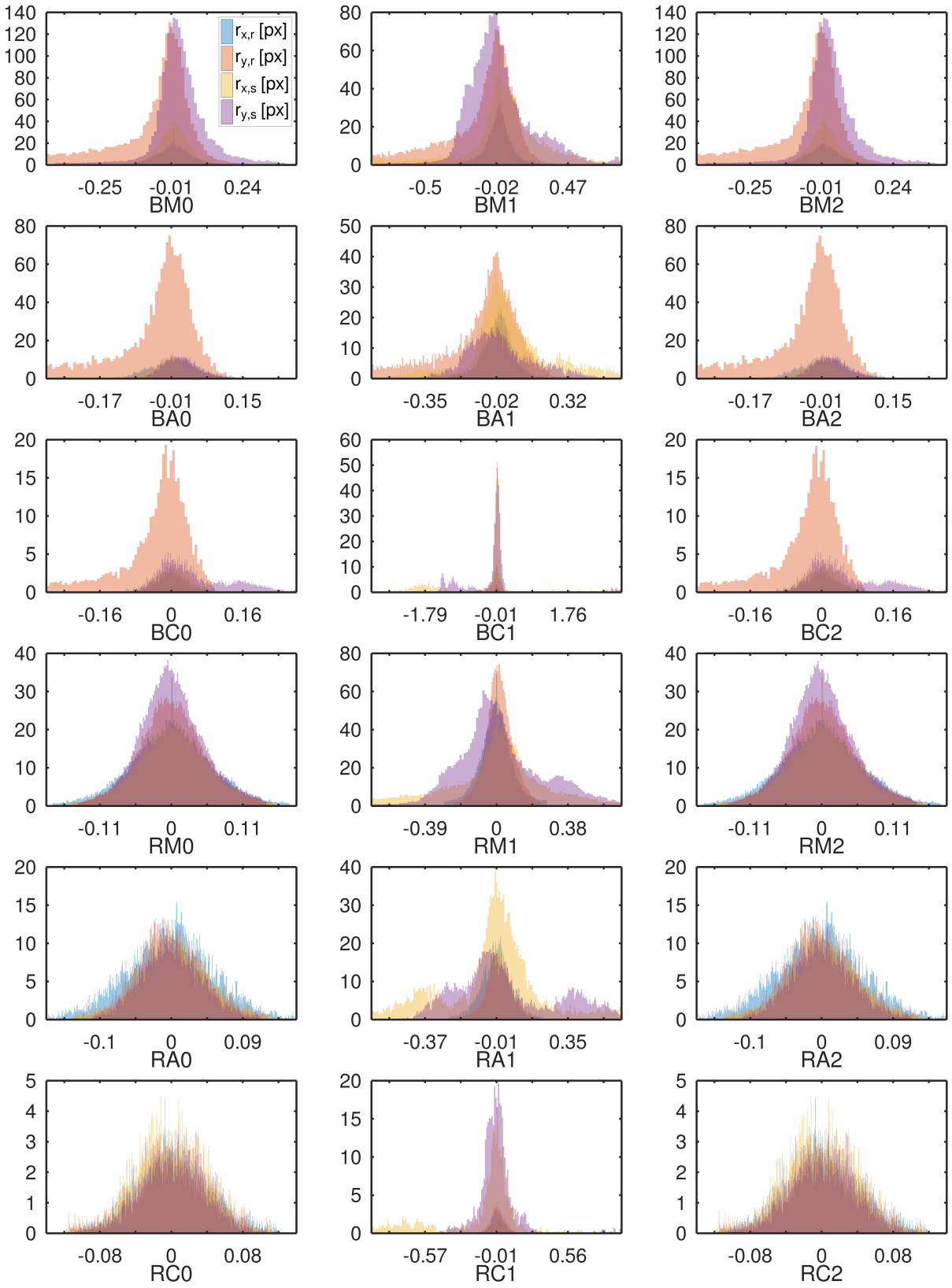


Figure 8.1: Histograms for the residuals r_x and r_y in x- and y-direction for both the reference (r) and slave (s) camera for the *without windshield* setup. Histograms given for all feature types (M: *Merged*, A: *Aicon*, C: *Checkerboard*), for all stereo constraints (*0*: no relative orientation, *1*: relative and exterior orientation of one camera, *2*: relative and exterior orientation of both cameras) and for two camera models (B: *Both*, R: *Radial*). Histograms normalized over parameter settings and experimental runs. Horizontal axis with unit [px] and limited to the $\pm 3.5\sigma$ interval.

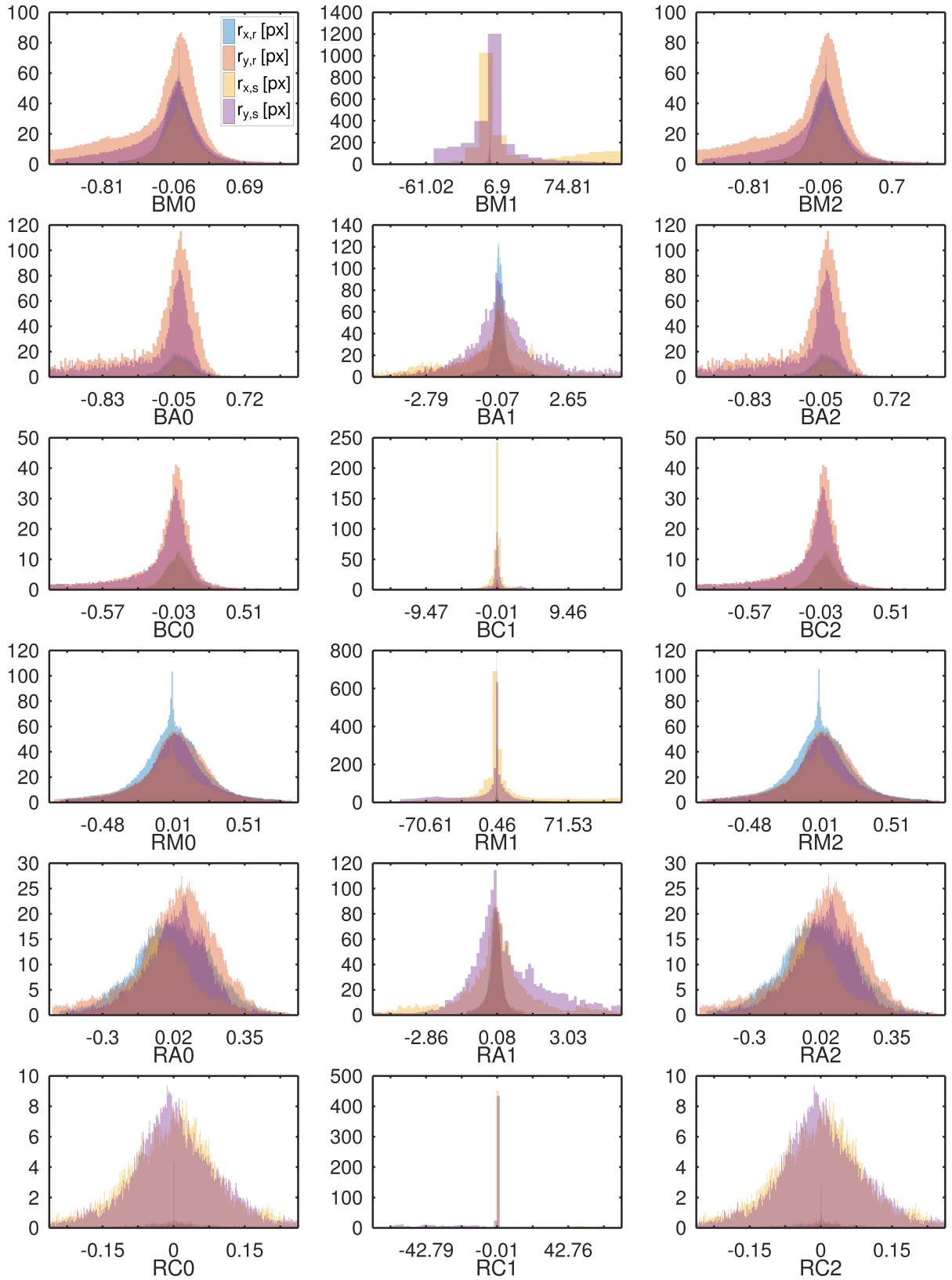


Figure 8.2: Histograms for the residuals r_x and r_y in x- and y-direction for both the reference and slave camera for the *with windshield* setup. Histograms given for all feature types, for all stereo constraints and for two camera models. Descriptions of the abbreviations see Figure 8.1. Histograms normalized over parameter settings and experimental runs. Horizontal axis with unit [px] and limited to the $\pm 3.5\sigma$ interval.

8.2.2 Deviations of orientation parameters between the *without* and *with windshield* setup

For each experimental case, relative deviations d of the estimated orientation parameter values and standard deviations between both setups (wwo) are calculated. For instance, d is calculated for the estimated parameter values ($\boldsymbol{X}$) of the interior orientation (I) as

$$\boldsymbol{d}_{wwo,\boldsymbol{X},I} = \frac{\hat{\boldsymbol{X}}_{I,wo} - \hat{\boldsymbol{X}}_{I,w}}{\hat{\boldsymbol{X}}_{I,wo}} \tag{8.1}$$

where $\hat{\boldsymbol{X}}_{I,wo}$ and $\hat{\boldsymbol{X}}_{I,w}$ are the vectors of estimated interior orientation parameter values in the *without windshield* (wo) and *with windshield* (w) setup, respectively. As the *without windshield*

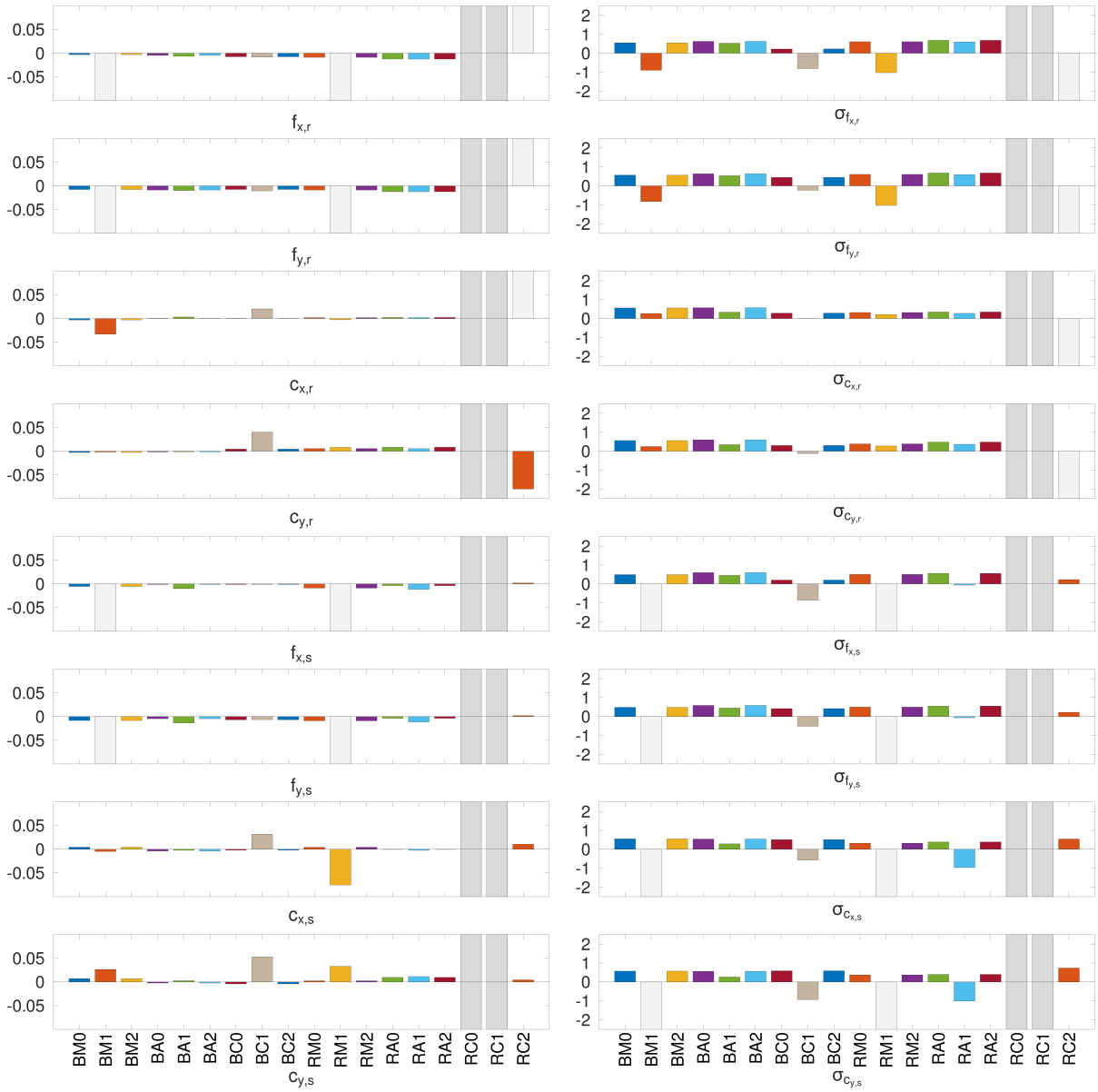


Figure 8.3: Relative deviations of focal length and principal point parameters between the *with* and *without windshield* setup. Reference and slave camera, all feature types, all stereo constraints and two camera models shown. Descriptions of the abbreviations see Figure 8.1. Vertical axis unit-free, invalid deviations in medium gray, off-limit deviations in light gray.

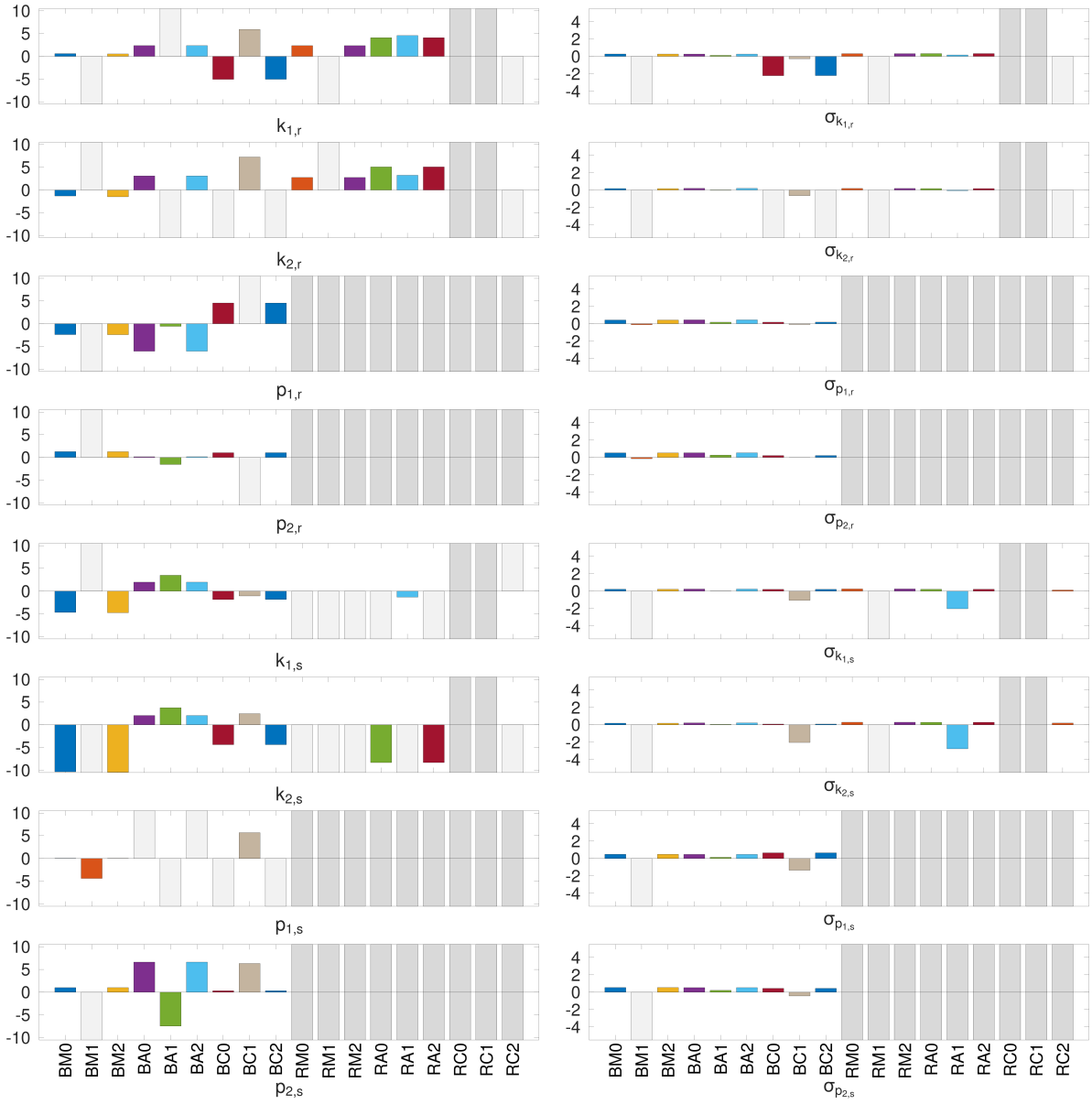


Figure 8.4: Relative deviations of radial and tangential distortion parameters between the *with* and *without windshield* setup. Reference and slave camera, all feature types, all stereo constraints and two camera models shown. Descriptions of the abbreviations see Figure 8.1. Vertical axis unit-free, invalid deviations in medium gray, off-limit deviations in light gray.

setup represents regular camera calibration and thus can be seen as kind of reference, it serves for normalization in the denominator. The deviations for the standard deviations $\boldsymbol{d}_{wwo,\sigma_{\boldsymbol{X}},I}$ of the interior orientation parameters, the deviations $\boldsymbol{d}_{wwo,\boldsymbol{X},R}$ and $\boldsymbol{d}_{wwo,\sigma_{\boldsymbol{X}},R}$ of the relative and the deviations $\boldsymbol{d}_{wwo,\sigma_{\boldsymbol{X}},E}$ of the exterior orientation parameters are obtained accordingly. Each deviation is averaged over all experimental runs and parameter settings. For the exterior orientation, the deviations are additionally averaged over all images; the three rotation angles and three position parameters are further averaged into a single parameter each time (σ_{rot} and σ_{pos}; cf. second paragraph in Section 8.2). Note that for $\boldsymbol{d}_{wwo,\boldsymbol{X}}$, an interpretation of positive and negative deviations as *better* or *worse* is not possible. For $\boldsymbol{d}_{wwo,\sigma_{\boldsymbol{X}}}$, positive deviations can be interpreted as that the standard deviations in the *without* setup are larger than in the *with* setup, hence the *with* setup has a better precision; for negative deviations, the opposite applies. Experimental

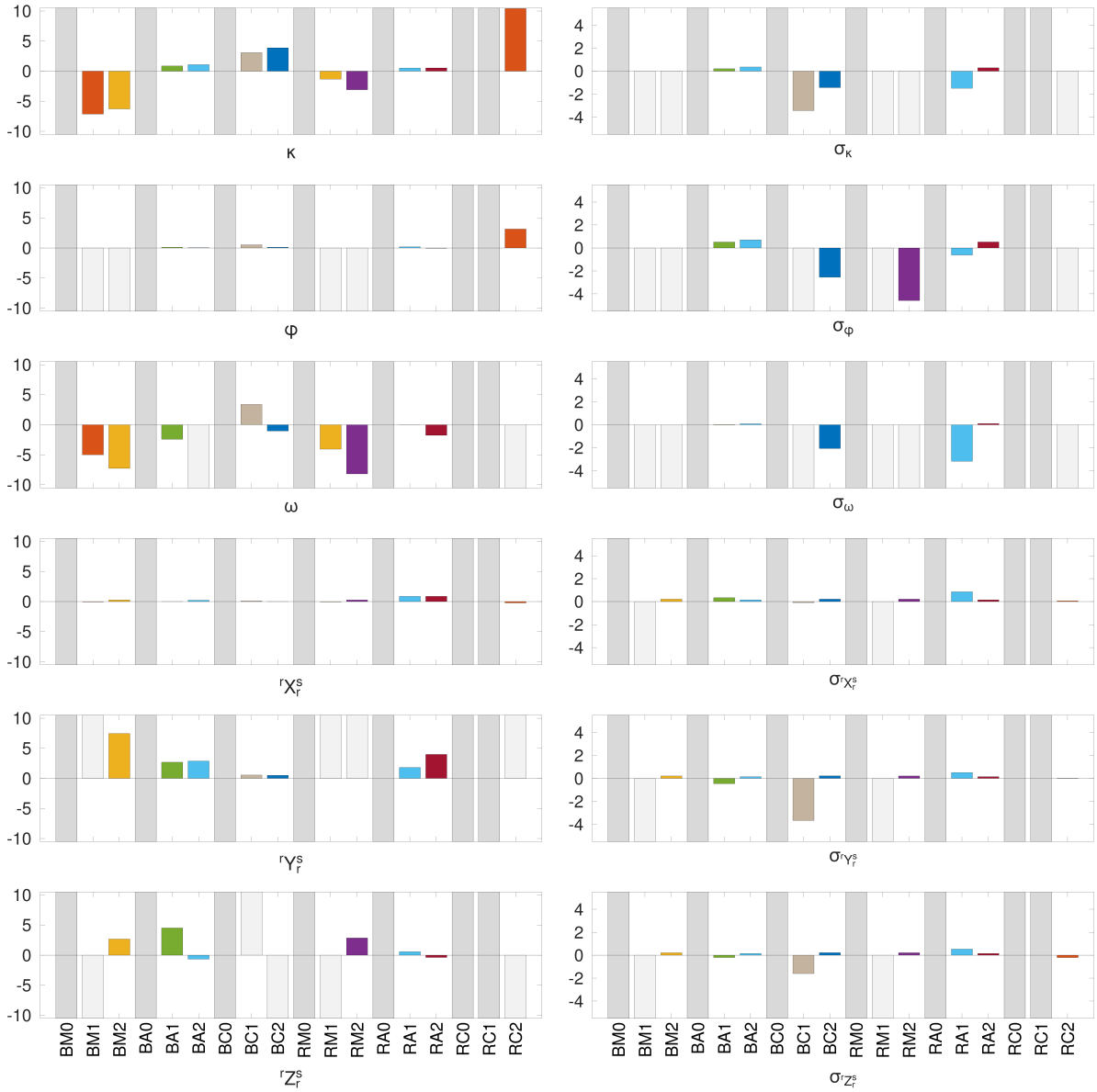


Figure 8.5: Relative deviations of relative orientation parameters between the *with* and *without windshield* setup. All feature types, all stereo constraints and two camera models shown. Descriptions of the abbreviations see Figure 8.1. Vertical axis unit-free, invalid deviations in medium gray, off-limit deviations in light gray.

cases with invalid values, i.e. where calibration was not successful or where parameters are not supported (e.g. tangential distortion for camera model *Radial*), are shown as medium gray bars in the deviation plots. The scaling of the vertical axis is chosen with regard to a trade-off between visibility and expressiveness of the plots. Such off-limit deviations exceeding the plots contain valid estimates and are displayed as light gray bars in the plots.

From the deviation plots of the interior orientation parameters, it can be seen that the deviations are in tendency larger for the distortion parameters (Figure 8.4) than for the focal length or principal point coordinates (Figure 8.3); note the different scaling of the vertical axes. The deviations for the standard deviations $\boldsymbol{d}_{wwo,\sigma_{\boldsymbol{X}},I}$ are in tendency larger than the deviations $\boldsymbol{d}_{wwo,\boldsymbol{X},I}$ for the parameter values. The deviations are also larger for *Stereo constraint 1* than for *0* and *2*. For camera model *Radial*, the deviations for the distortion parameters are in tendency larger

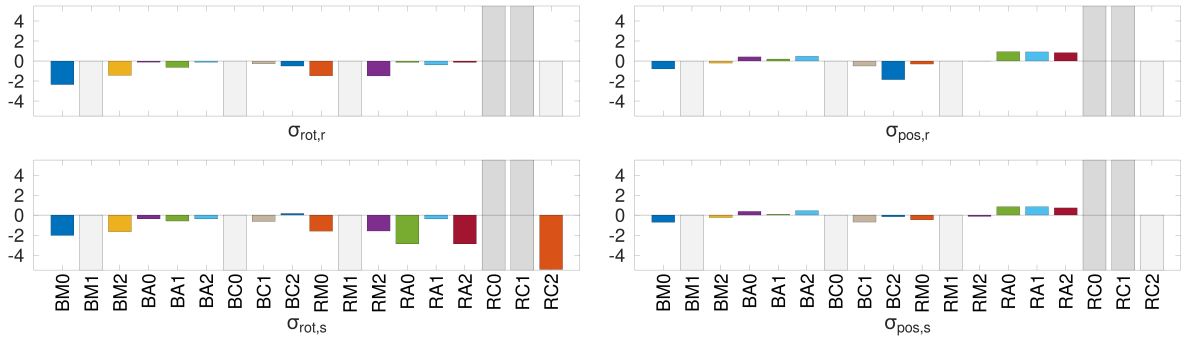


Figure 8.6: Relative deviations of averaged standard deviations of exterior orientation parameters between the *with* and *without windshield* setup. Reference and slave camera, all feature types, all stereo constraints and two camera models shown. Descriptions of the abbreviations see Figure 8.1. Vertical axis unit-free, invalid deviations in medium gray, off-limit deviations in light gray.

than for camera model *Both*, what indicates that the images could contain tangential distortions that are not modeled by *Radial*. Furthermore, for feature type *Checkerboard* and camera model *Radial*, calibration was not successful for two experimental cases. In comparison between the reference and slave camera, there are only a few parameters with visually larger deviations for one camera (e.g. $d_{wwo,\boldsymbol{X},f_x}$ large for r, small for s). From the deviation plots of the relative orientation parameters (Figure 8.5; angles given as Euler angles) it can be observed that in tendency $\boldsymbol{d}_{wwo,\boldsymbol{X},R}$ are larger than $\boldsymbol{d}_{wwo,\sigma_{\boldsymbol{X}},R}$. In general, the differences between the experimental cases are large (many off-limit deviations, light gray color). $\boldsymbol{d}_{wwo,\boldsymbol{X},R}$ are larger for *Stereo constraint 2* than for *1*, which seems reasonable as the definition of the constraints in *1* avoids contradictions. For $\boldsymbol{d}_{wwo,\sigma_{\boldsymbol{X}},R}$, the observation is the opposite. Furthermore, the deviations are larger for feature type *Merged* than for *Aicon*. From the deviation plots of the exterior orientation parameters (Figure 8.6), the only remarkable observation that can be made is that $\boldsymbol{d}_{wwo,\sigma_{\boldsymbol{X}},E}$ is larger for the rotation than for the position parameters. Finally, for all three kinds of orientation parameters, it can be observed that the precision is smaller in average for the *without windshield* setup.

8.2.3 Significance tests on deviations between the *without* and *with windshield* setup

In contrast to visual observations made from plots, statistical tests (Table 8.2) do not depend on subjective perception and thus are considered as complementary aspect for analysis of the influence of the windshield. The tests are performed for the estimated parameter values and standard deviations of the interior, relative and exterior orientation and evaluate whether there is a statistically significant difference between the *with* and *without* setup. The selected significance level is $\alpha = 0.05$ for all tests. For the estimated parameter values, a two-tailed t-test with the null hypothesis H_0 and alternative hypothesis H_1 is performed. The hypotheses are defined as shown exemplarily for the interior orientation by

$$H_0 : \hat{\boldsymbol{X}}_{I,wo} = \hat{\boldsymbol{X}}_{I,w} \tag{8.2}$$

$$H_1 : \hat{\boldsymbol{X}}_{I,wo} \neq \hat{\boldsymbol{X}}_{I,w} \tag{8.3}$$

whereby the mathematical items are defined as in Subsection 8.2.2. As the variances and redundancies of $\hat{\boldsymbol{X}}_{I,wo}$ and $\hat{\boldsymbol{X}}_{I,w}$ are not equal, the hypothesis tests are carried out as Welch's t-test

[Welch, 1947]. For the estimated standard deviations, a two-tailed F-test is performed. The hypotheses are defined as shown exemplarily for the interior orientation by

$$H_0 : \hat{\boldsymbol{\sigma}}^2_{\hat{\boldsymbol{X}}_{I,wo}} = \hat{\boldsymbol{\sigma}}^2_{\hat{\boldsymbol{X}}_{I,w}} \tag{8.4}$$

$$H_1 : \hat{\boldsymbol{\sigma}}^2_{\hat{\boldsymbol{X}}_{I,wo}} \neq \hat{\boldsymbol{\sigma}}^2_{\hat{\boldsymbol{X}}_{I,w}} \tag{8.5}$$

whereby $\hat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{X}}_{I,wo}}$ and $\hat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{X}}_{I,w}}$ are the vectors of the estimated standard deviations of the interior orientation parameters in the *without windshield* (wo) and *with windshield* (w) setup, respectively.

Table 8.2: Results of significance tests between the *with* and *without windshield* setup for the interior, relative and exterior orientation parameters for both cameras, all feature types, for all stereo constraints and for two camera models. Invalid deviations in medium gray background color. Descriptions of the abbreviations see Figure 8.1. Non-significant deviation for value 0, significant deviation otherwise.

	BM0	BM1	BM2	BA0	BA1	BA2	BC0	BC1	BC2	RM0	RM1	RM2	RA0	RA1	RA2	RC0	RC1	RC2
$f_{x,r}$																		
$f_{y,r}$																		
$c_{x,r}$				0			0		0									
$c_{y,r}$							0											
$k_{1,r}$					0		0											
$k_{2,r}$							0											
$p_{1,r}$							0											
$p_{2,r}$				0			0											
$f_{x,s}$							0	0										
$f_{y,s}$																		
$c_{x,s}$							0		0				0					
$c_{y,s}$							0											
$k_{1,s}$								0						0				
$k_{2,s}$								0						0				
$p_{1,s}$																		
$p_{2,s}$							0											
κ																		
φ						0			0						0			
ω																		
${}^rX^s_r$																		
${}^rY^s_r$						0			0			0			0			0
${}^rZ^s_r$						0						0			0			
$\sigma_{f_{x,r}}$								0										0
$\sigma_{f_{y,r}}$								0										0
$\sigma_{c_{x,r}}$								0										0
$\sigma_{c_{y,r}}$								0										0
$\sigma_{k_{1,r}}$							0	0	0									0
$\sigma_{k_{2,r}}$					0		0	0	0					0				0
$\sigma_{p_{1,r}}$								0										
$\sigma_{p_{2,r}}$								0										
$\sigma_{f_{x,s}}$								0						0				
$\sigma_{f_{y,s}}$								0						0				
$\sigma_{c_{x,s}}$								0						0				
$\sigma_{c_{y,s}}$								0						0				
$\sigma_{k_{1,s}}$					0			0						0				
$\sigma_{k_{2,s}}$					0		0	0	0					0				
$\sigma_{p_{1,s}}$								0										
$\sigma_{p_{2,s}}$								0										
σ_κ								0	0			0		0				0
σ_φ								0	0			0		0				0
σ_ω					0			0	0			0		0				0
$\sigma_{{}^rX^s_r}$								0										
$\sigma_{{}^rY^s_r}$					0			0						0				
$\sigma_{{}^rZ^s_r}$					0			0						0				0
$\sigma_{rot,r}$				0	0	0	0	0	0	0		0		0				0
$\sigma_{pos,r}$							0	0	0	0		0						0
$\sigma_{rot,s}$				0	0	0	0	0	0	0		0	0	0	0			0
$\sigma_{pos,s}$					0		0	0		0		0						0

Looking at the test results for the parameter values, the rate of statistically non-significant deviations is highest for the principal point parameters, followed by the tangential distortions, then by the radial distortions. The rate is lowest for the focal length parameters. For the position parameters of the relative orientation, the rate is higher than for the rotation parameters. Looking at the three feature types, *Checkerboard* has a higher rate of non-significant differences than *Aicon* than *Merged*. Comparing the stereo constraints with each other, the rate of non-significant deviations is highest for *0*, followed by *2*, while it is lowest for *1*. Likewise, the rate of non-significant differences is higher for camera model *Both* than for *Radial*. For neither of these parameters and for neither of these experimental cases, more than one quarter of all valid deviations is statistically non-significant, i.e. there is a statistical significance of the deviations between the *with* and *without* setup in the majority of experimental cases.

Looking at the test results for the standard deviations, the rate of non-significant deviations is higher for k than for f and c, which are equal, than for p. For the relative orientation, the rate is higher for the position than for the rotation parameters, which is the opposite from the tests on parameter values. In contrast, for the exterior orientation, the rate is higher for the rotation parameters. For the three stereo constraints, now *1* shows the highest rate of non-significant deviations, followed by *2*, then by *0*. The observations for the three feature types and the two camera models are coherent with the tests on the deviations of the parameter values. Finally, in contrast to the deviations on the parameter values, there are several parameters and experimental cases where the rate of non-significant deviations is above one quarter. With feature type *Checkerboard*, there is also one experimental case, where the majority of deviations is statistically non-significant.

8.2.4 Correlations between orientation parameters

As large correlations between the estimated camera calibration parameters are not desired (cf. Chapter 4), their values are analyzed in the following. Therefore, the correlations are plotted for the *Both* (Figures 8.7, 8.9) and for the *Radial* (Figures 8.8, 8.10) camera model for each setup. Hereby, for the sake of a meaningful visualization, the correlations are put into six correlation groups (Table 8.3). Each group covers the correlations of all contributing orientation parameters obtained by calibration with all parameter settings and experimental runs for the respective camera model and setup. For instance, group I covers the interior orientation parameters, i.e. the focal length, the principal point, the radial-symmetric distortion and the tangential distortion parameters. The analysis addresses remarkable observations about the mean correlation, the standard deviation and the value range of the correlations in each group.

Table 8.3: Correlation groups defined to evaluate the correlations among and between orientation parameters for camera calibration with test fields.

I: Among interior orientation	IR: Between interior and relative orientation
R: Among relative orientation	IE: Between interior and exterior orientation
E: Among exterior orientation	RE: Between relative and exterior orientation

The mean correlations (plus sign in the four correlation figures) are between ± 0.3 for all experimental cases and correlation groups for both setups. These small values indicate the absence of systematic deficiencies in the imaging configuration. Remarkable differences in the mean correlations between the experimental cases are not observed in the figures. Analysis of the standard deviations of the correlations (1σ interval visualized by the black horizontal lines in the four correlation figures) can give insights on how strong the correlations differ between the orientation parameters, experimental runs and parameter settings within each group. For instance, a small standard deviation indicates that the correlations are similar for all parameters, parameter set-

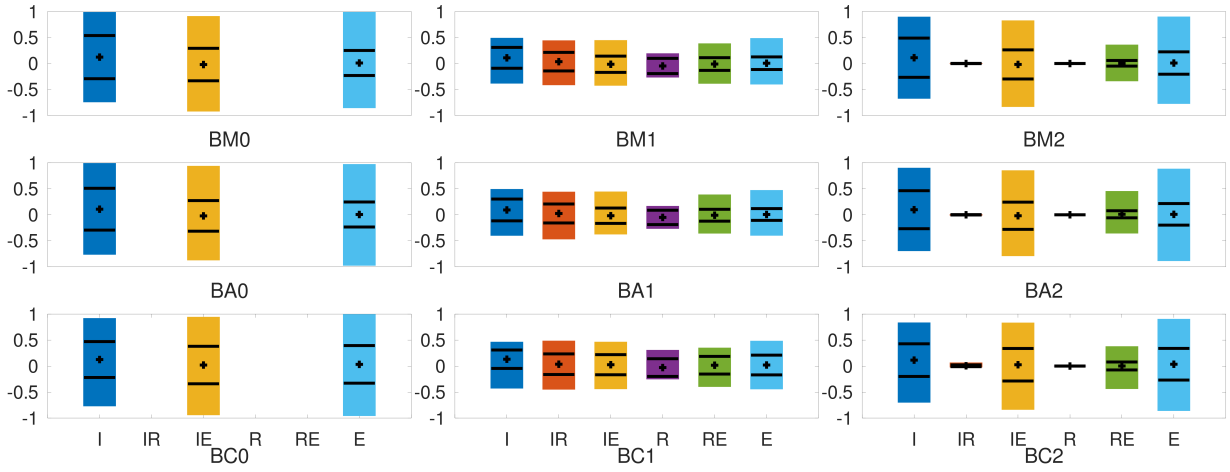


Figure 8.7: Correlations among the interior (I), relative (R) and exterior (E) orientation parameters and correlations between all pair-wise combinations (IR, IE, RE) for the *Both* camera model and the *without windshield* setup for all feature types and all stereo constraints. The colored bars show the value range from minimal to maximal correlation, the plus signs show the mean correlation and the two black lines show the 1σ interval. Descriptions of the remaining abbreviations see Figure 8.1. Vertical axis normalized to $[-1, 1]$.

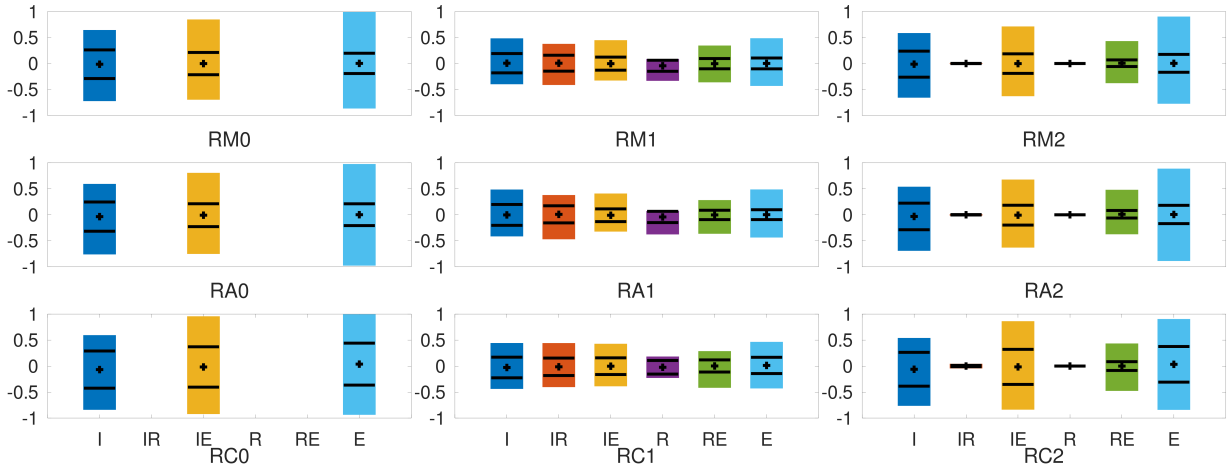


Figure 8.8: Correlations among the interior, relative and exterior orientation parameters and correlations between all pair-wise combinations for the *Radial* camera model and the *without windshield* setup for all feature types and all stereo constraints. The colored bars show the value range from minimal to maximal correlation, the plus signs show the mean correlation and the two black lines show the 1σ interval. Descriptions of the abbreviations see Figures 8.1 and 8.7. Vertical axis normalized to $[-1, 1]$.

tings and runs. In the following, one group after the other is discussed. Looking at group *I*, the standard deviations are in tendency smaller for *1* than for *0* and *2*. They are visually remarkably smaller for one camera model for some experimental cases and larger for others. The same applies with regard to the three feature types. Especially cases with feature type *Both* show larger standard deviations for the *without* setup. Looking at group *IR*, the standard deviations are in tendency smaller for *2* than for *1* and smaller for the *with* than for *without* windshield setup. Looking at group *IE*, standard deviations are in tendency smaller for *1* than for *2* than for *0*. They are smaller for *Radial* than for *Both*. They are visually larger for *Checkerboard*, especially with *Radial*. With regard to the setups, the standard deviations are larger for *without*. Looking at group *R*, the same observation can be made for the stereo constraints as for *IR*. In tendency, the standard deviations are larger for *Both* than for *Radial*, except for one experimental case, and

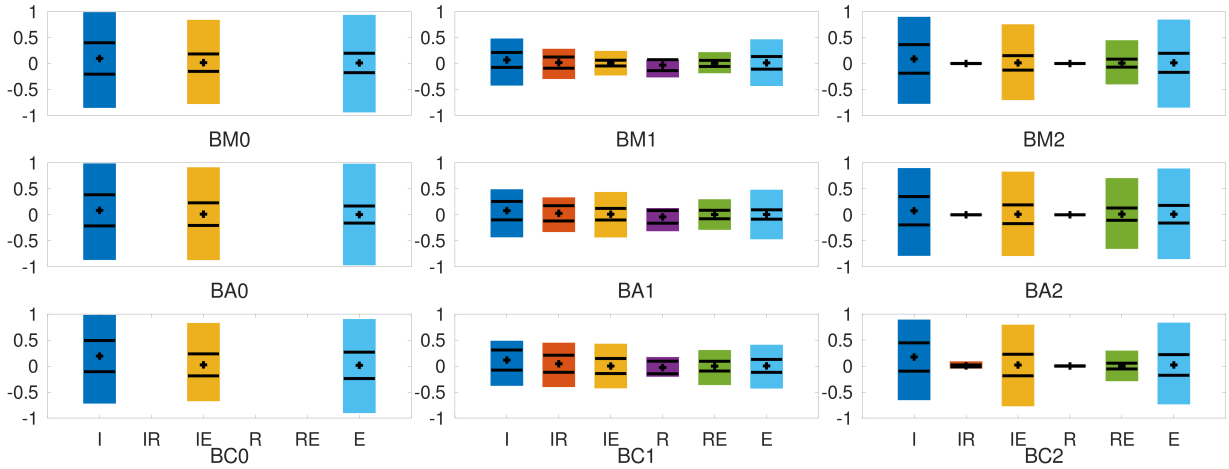


Figure 8.9: Correlations among the interior, relative and exterior orientation parameters and correlations between all pair-wise combinations for the *Both* camera model and the *with windshield* setup for all feature types and all stereo constraints. The colored bars show the value range from minimal to maximal correlation, the plus signs show the mean correlation and the two black lines show the 1σ interval. Descriptions of the abbreviations see Figures 8.1 and 8.7. Vertical axis normalized to $[-1, 1]$.

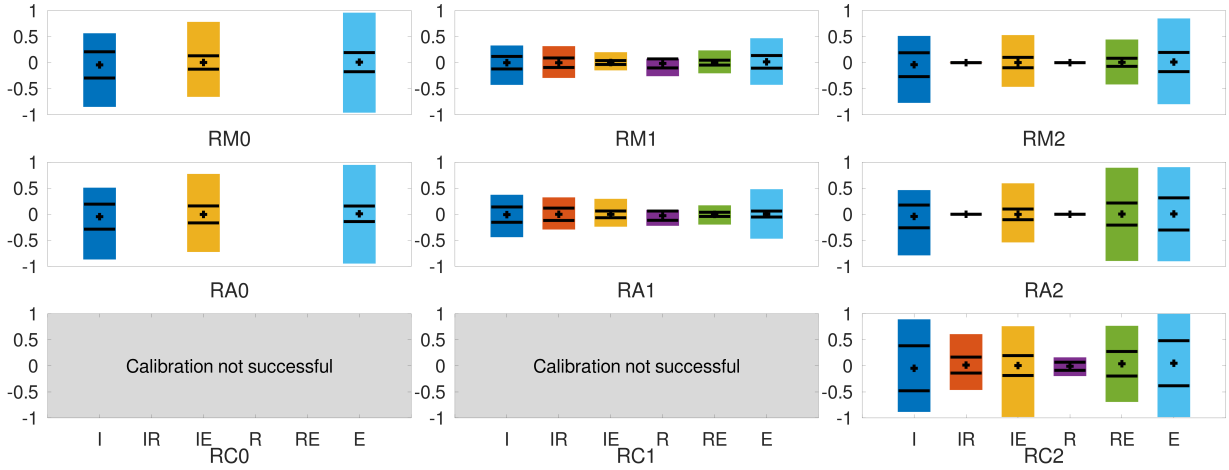


Figure 8.10: Correlations among the interior, relative and exterior orientation parameters and correlations between all pair-wise combinations for the *Radial* camera model and the *with windshield* setup for all feature types and all stereo constraints. The colored bars show the value range from minimal to maximal correlation, the plus signs show the mean correlation and the two black lines show the 1σ interval. Descriptions of the abbreviations see Figures 8.1 and 8.7. Vertical axis normalized to $[-1, 1]$.

they are larger for *Checkerboard* than for *Aicon* and *Merged*. The standard deviations are larger for the *without* than for *with* windshield setup. Looking at group *RE*, *Checkerboard* and *Aicon*, respectively, show visually larger standard deviations than the other two feature types for a few experimental cases each, while *Merged* doesn't. Looking at group *E*, the standard deviations are in tendency smaller for *1* than for *0* and *2*; comparing the later two, the standard deviations are sometimes smaller for the one, sometimes for the other. With regard to the feature types and camera models, *Checkerboard* shows larger standard deviations especially with camera model *Radial*.

Analysis of the value ranges (minimal and maximal correlation values visualized by the lower and upper end of the colored bars in the four correlation plots) can reveal additional insights compared to the standard deviations, which only cover approx. 65% of all contributing correlation values. For example, if the value range is not symmetric around the standard deviation in these

figures, it may be an indicator for outliers or systematic effects in the correlations, similar to the skew measure for histograms. With regard to the stereo constraints, deviations from symmetry are observed in tendency more often and stronger for *1* than for *0* or *2*. The same applies for *Merged* and *Aicon* than for *Checkerboard*. For the *without* setup, such deviations are also observed more often and stronger than for the *with* setup. Most deviations from symmetry occur for *IE*, *R* and *RE*. Nevertheless, the mentioned deviations are visually interpreted smaller than 1σ, except for group *R*, where the minimal correlation value is far below the lower end of the $\pm 1\sigma$ interval.

8.2.5 Discussion

Several aspects need to be acknowledged with regard to the calibration setup that may influence the comparability between the two setups or that could be a potential improvement, but at the cost of a higher effort or a calibration setup that is practically less feasible. First, the advantage of the virtual test field has not been fully exploited: The 2d test fields have been placed close to each other in all images, but they could have been placed more distant from each other for some images, or a larger number of test fields could have been used to achieve a better geometric configuration [cf. Geiger et al., 2012]. Additionally, large-scale reference marks leading to a better detectability at distant test field positions could have been used to increase depth coverage. Second, though primary objective of image acquisition was to obtain a valid imaging configuration for calibration [cf. Luhmann et al., 2006] for both setups, it was neither ensured that the number of images nor the camera positions and orientations are exactly the same in the two setups. Furthermore, an available spare windshield would have allowed to acquire the images for the *with* setup in the same (laboratory) conditions as for the *without* setup, which has been done under different conditions in a real car.

The used algorithm can be seen as quite complex for test field-based camera calibration and could have been simplified at several steps, as the following aspects indicate. But as the calibration parameters are estimated in a final bundle adjustment, an influence of the complex algorithm in the preceding steps on the estimated parameters is not expected and the simplifications would just be beneficial with regard to computational power or implementation effort. First, the use of coded marks only could have facilitated feature matching and object coordinate association by easy assignment based on their point numbers and could have allowed to get rid of structure-from-motion as no temporary 3d object coordinates would have been required. Nevertheless, the complexity of the algorithm seems not to be exceptional compared to other calibration frameworks, as the user interface and manual of the software "Aicon 3D Studio" with equal functionality let assume. It appears that Aicon uses algorithms with similar complexity to handle point association of uncoded marks between different images [Schneider et al., 2017]; additionally, as far as known, this software does not support the automatic integration of pre-determined 3d coordinates for uncoded marks, only for coded marks. Furthermore, own experiences made with this software in previous work [Hanel et al., 2016] with the same dataset has shown that automatic assignment for uncoded marks was wrong in many cases so that several images could not be registered and therefore a lot of manual correction was required. In the presented evaluation, problems with point association can be seen only at a minor extent in the *with* setup (higher number of estimated 3d points than reference marks available on the test fields in cases *RA* and *OC* in Table 8.1). Second, the use of test fields at fixed positions [cf. Geiger et al., 2012] could have spared the complex technical implementation of the virtual 3d test field.

Several aspects should also be acknowledged with regard to the evaluation. First, a Kolmogorov-Smirnov test [Massey, 1951] for normality of the residuals has been done exemplarily for several experimental cases, but all of them failed, i.e. the distributions are seen as not normally-distributed. As such analytical tests may discard normal distributions for large sample sizes - as present - for

small deviations from normality already, the relevance of these test results should not be overestimated [Ghasemi & Zahediasl, 2012]: Visual analysis of distributions, e.g. by histograms or quantile-quantile plots (Q-Q plots) can be seen as complementary alternative to analytical tests [Ghasemi & Zahediasl, 2012]. As histograms allow the reader to inspect the distributions himself, they have been chosen for visualization of the residuals (cf. Subsection 8.2.1). Additional Q-Q plots, which are not shown in this thesis, have been created exemplarily for some experimental cases assuming normal distribution one time and Laplacian distribution the other time. These plots indicate a better fit (i.e. points closer to the identity line) of the residuals to the Laplacian distribution than to the normal distribution. As this observation could be an indicator for the presence of outliers, it motivated the implementation of a robust loss (cf. parameter settings) for the cost function for least-squares optimization, which allows to handle outliers (that would cause deviations from normal distribution) better than a regular cost function [Triggs et al., 2000; Agarwal et al., 2022]. Second, only the evaluation metrics considered as most important for answering the research questions are presented in this thesis. There are a plenty of other common measures to evaluate camera calibration results like the number and average intersection angle of image rays per point [Luhmann et al., 2006], which have been not considered. Third, for the desirable assessment of the accuracy, no independent reference information was available. Though some points on the test fields could have been excluded from calibration and so serve as analytically independent reference information, they would still have been affected by the same hardware-sided systematic effects (e.g. wrong metric scale) as other points on the same test field and hence would not be entirely independent. Therefore, a completely independent reference object (e.g. a third test field) with object coordinates determined with better accuracy than by photogrammetry used for the proposed work (see Subsection 4.2.1) could be a solution to provide completely independent reference information.

8.3 Camera calibration with traffic signs

For camera calibration with traffic signs, the evaluation comprises statistical measures, deviation plots and hypothesis tests. All shown values are averages from 10 experimental runs.

8.3.1 Statistics

The following six statistical measures are presented: The number of image points and object points of the reference points as well as the mean values and standard deviations of the residuals of the image points in x- and y-direction. In contrast to the evaluation of camera calibration with test fields (cf. Section 8.2), the residuals are described by numeric values assuming normal distribution as no disturbing factors like a vehicle windshield play a role in this method.

Clearly more reference points for calibration can be extracted from the *Munich sequence* (Table 8.5) in contrast to the *Ettlingen sequence* (Table 8.4). For some experimental cases (e.g. Ettlingen, 1 R), the number of reference points is even smaller than it can be expected from a single image of a typical calibration test field. In such cases, it is questionable whether the calibration parameters can be estimated reliably or whether outliers in the observations can be detected. Looking at both sequences, the largest mean residuals have exponent -8. Based on this small value and in comparison to a typical image measurement accuracy with an exponent of -1 (i.e. 0.1 px [Luhmann et al., 2016]), no bias in the pixel coordinates of the reference points can be identified. For the x- and y-direction in either of the two sequences, the minimal and maximal exponents of the standard deviations are equal up to differences of ± 1. Therefore, systematic scattering of the residuals that is stronger in one of the two directions does not become visible.

Table 8.4: Statistical measures for calibration with traffic signs for the *Ettlingen sequence* for all experimental cases (C: circular traffic sign shape, R: rectangular shape, T: triangular shape; 1,2,3: different semantic segmentation and depth estimation methods). $\bar{r}$ represents the mean residuals of the image points in x- and y-direction and σ_r represents the standard deviation of the residuals in x- and y-direction.

Experimental case	#image points	#object points	$\bar{r}_x$ [px]	$\bar{r}_y$ [px]	σ_{r_x} [px]	σ_{r_y} [px]
1 C	13	13	$-1.50 \cdot 10^{-08}$	$4.84 \cdot 10^{-08}$	$9.11 \cdot 10^{-07}$	$5.22 \cdot 10^{-07}$
1 R	8	8	$-3.25 \cdot 10^{-08}$	$-9.49 \cdot 10^{-08}$	$8.64 \cdot 10^{-08}$	$2.24 \cdot 10^{-07}$
1 T	51	51	$-2.84 \cdot 10^{-10}$	$5.87 \cdot 10^{-10}$	$4.93 \cdot 10^{-08}$	$9.59 \cdot 10^{-09}$
1 CR	21	21	$-1.61 \cdot 10^{-08}$	$8.56 \cdot 10^{-09}$	$1.23 \cdot 10^{-07}$	$1.49 \cdot 10^{-07}$
1 CT	64	64	$1.36 \cdot 10^{-08}$	$-8.22 \cdot 10^{-09}$	$2.04 \cdot 10^{-07}$	$2.71 \cdot 10^{-07}$
1 RT	59	59	$9.93 \cdot 10^{-10}$	$-2.72 \cdot 10^{-09}$	$1.56 \cdot 10^{-08}$	$7.70 \cdot 10^{-08}$
1 CRT	72	72	$-2.16 \cdot 10^{-09}$	$1.08 \cdot 10^{-09}$	$1.13 \cdot 10^{-07}$	$5.87 \cdot 10^{-08}$
2 CRT	72	72	$1.46 \cdot 10^{-10}$	$1.01 \cdot 10^{-10}$	$5.18 \cdot 10^{-09}$	$1.82 \cdot 10^{-08}$
3 CRT	66	66	$-7.05 \cdot 10^{-09}$	$1.18 \cdot 10^{-09}$	$7.80 \cdot 10^{-07}$	$1.53 \cdot 10^{-07}$

Table 8.5: Statistical measures for calibration with traffic signs for the *Munich sequence* for all experimental cases (C: circular traffic sign shape, R: rectangular shape, T: triangular shape; 1,2,3: different semantic segmentation and depth estimation methods). $\bar{r}$ represents the mean residuals of the image points in x- and y-direction and σ_r represents the standard deviation of the residuals in x- and y-direction.

Experimental case	#image points	#object points	$\bar{r}_x$ [px]	$\bar{r}_y$ [px]	σ_{r_x} [px]	σ_{r_y} [px]
1 C	68	68	$2.91 \cdot 10^{-08}$	$2.14 \cdot 10^{-08}$	$2.80 \cdot 10^{-07}$	$2.11 \cdot 10^{-07}$
1 R	78	78	$1.06 \cdot 10^{-07}$	$4.49 \cdot 10^{-09}$	$6.81 \cdot 10^{-07}$	$2.38 \cdot 10^{-08}$
1 T	82	82	$1.76 \cdot 10^{-11}$	$1.02 \cdot 10^{-11}$	$1.27 \cdot 10^{-10}$	$7.15 \cdot 10^{-11}$
1 CR	147	147	$-2.53 \cdot 10^{-07}$	$-1.23 \cdot 10^{-07}$	$3.46 \cdot 10^{-06}$	$2.09 \cdot 10^{-06}$
1 CT	149	149	$-5.11 \cdot 10^{-09}$	$-1.51 \cdot 10^{-08}$	$1.40 \cdot 10^{-06}$	$4.47 \cdot 10^{-07}$
1 RT	159	159	$3.86 \cdot 10^{-09}$	$5.53 \cdot 10^{-09}$	$4.03 \cdot 10^{-07}$	$2.74 \cdot 10^{-07}$
1 CRT	227	227	$1.20 \cdot 10^{-07}$	$-3.15 \cdot 10^{-07}$	$7.23 \cdot 10^{-06}$	$8.78 \cdot 10^{-06}$
2 CRT	228	228	$-7.80 \cdot 10^{-11}$	$-9.05 \cdot 10^{-11}$	$5.23 \cdot 10^{-09}$	$1.24 \cdot 10^{-08}$
3 CRT	26	26	$-1.62 \cdot 10^{-06}$	$-1.54 \cdot 10^{-06}$	$1.57 \cdot 10^{-06}$	$2.98 \cdot 10^{-06}$

8.3.2 Deviations of orientation parameters between the proposed and a reference calibration

The relative deviations $\boldsymbol{d}_{cr,\boldsymbol{X}}$ (Figure 8.11) for the estimated interior orientation parameter values are calculated between estimates from all experimental cases (index c) and the reference calibration (index r) according to

$$\boldsymbol{d}_{cr,\boldsymbol{X}} = \frac{\hat{\boldsymbol{X}}_{I,c} - \hat{\boldsymbol{X}}_{I,r}}{|\hat{\boldsymbol{X}}_{I,r}|} \tag{8.6}$$

with $\hat{\boldsymbol{X}}_{I,c}$ being the vector of estimated interior orientation parameter values in the experimental case c and $\hat{\boldsymbol{X}}_{I,r}$ being the vector of interior orientation parameter values from reference calibration. The formula for the relative deviations of standard deviations $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ is defined analogue. According to these formulae, a deviation of $+1$ can be interpreted as that the estimated value is twice as large as the reference value, while a deviation of -0.5 can be interpreted as that the estimated value is half of the reference value. For further details, see Section 8.1. Note the scaling of the vertical axis in the deviation plots has been selected so that differences between the experimental cases become clearly visible. Therefore, it became necessary to truncate extraordinary large deviations, why the corresponding bars of such off-limit deviations are set to a light gray color.

As tendency, in experimental cases with one traffic sign shape, $\boldsymbol{d}_{cr,\boldsymbol{X}}$ and $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ are smaller for most interior orientation parameters when triangles (T) are used than when rectangles (R) or

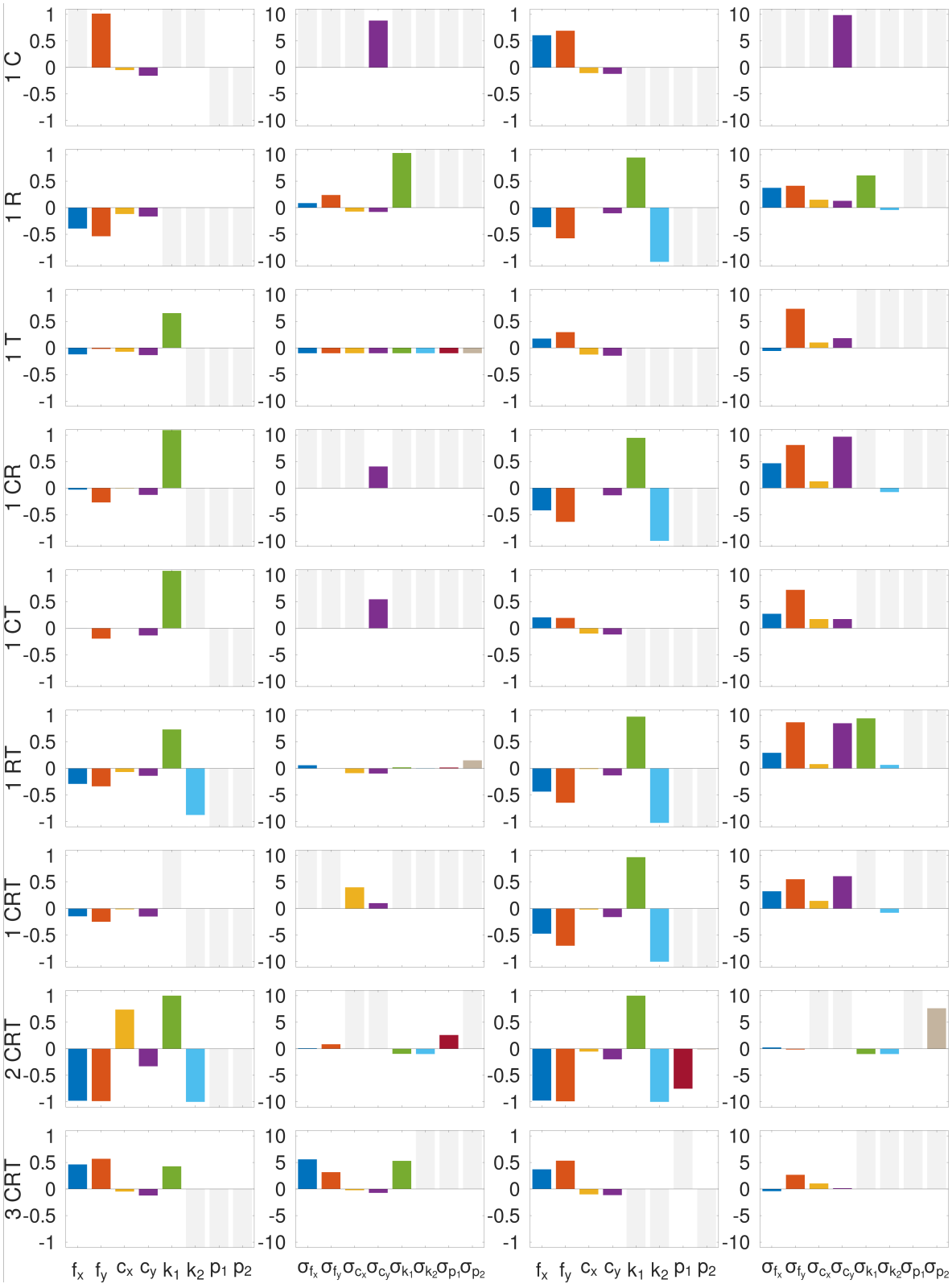


Figure 8.11: Relative deviations of the estimated interior orientation parameter values and standard deviations between camera calibration with traffic signs and reference calibration. The vertical axis is unit-free due to relative deviations, off-limit deviations are colored in light gray. Descriptions of the abbreviations see Table 8.4.

circles (C) are used. In experimental cases with two shapes, $\boldsymbol{d}_{cr,\boldsymbol{X}}$ are smaller for CT than for CR and RT. $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ are smallest for CT for the *Munich* and smallest for RT for the *Ettlingen sequence*. As it can be seen, smallest deviations are often obtained for experimental cases with triangle-shaped traffic signs. With all three shapes used, $\boldsymbol{d}_{cr,\boldsymbol{X}}$ is smallest for 1, while $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ is smallest for 2. 1 and 2 share with Deeplabv3+ the same semantic segmentation method, while 3 relies on EfficientPS. Hence, it seems favorable to use Deeplabv3+. As small deviations in the estimated parameters values are seen as more important than negative or small positive deviations in the estimated standard deviations ($\rightarrow$ limited expressiveness of precision, cf. Section 8.1), it has been decided after preliminary experiments to perform the remaining experimental cases with one or two shapes for 1, i.e. by using Monodepth2 as depth estimation and Deeplabv3+ as semantic segmentation method. But as no further conclusions are expected, the experimental cases with one or two shapes have not been performed for 2 or 3. In general, only a few negative values can be observed for $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$, i.e. cases where the proposed method has better precision than the reference calibration. Comparing the interior orientation and distortion parameters with each other, the deviations are often lower for c_x and c_y than for f_x and f_y and lower for k_1, k_2 than for p_1, p_2. The deviations for distortion parameters are often larger than for the interior orientation parameters. Furthermore, $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ are often larger than $\boldsymbol{d}_{cr,\boldsymbol{X}}$ and the deviations are larger for the *Ettlingen* than for the *Munich sequence*. Hence, there might be some degree of dependency on the type of road scene. Finally, remarkable correlations between the number of image points in the different experimental cases and the deviations do not become visible from the deviation plots. Likewise, correlations between $\boldsymbol{d}_{cr,\boldsymbol{X}}$ and $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ do not become visible.

8.3.3 Significance tests on deviations between estimated and reference orientation values

Intention of the hypothesis tests is to determine whether the deviations between the estimated and the reference parameter values (cf. Subsection 8.3.2) are statistically significant or not. The tests for parameter values are carried out with null hypothesis H_0 and alternative hypothesis H_1 defined as

$$H_0 : \hat{\boldsymbol{X}}_{I,c} = \hat{\boldsymbol{X}}_{I,r} \tag{8.7}$$

$$H_1 : \hat{\boldsymbol{X}}_{I,c} \neq \hat{\boldsymbol{X}}_{I,r} \tag{8.8}$$

whereby the mathematical items are defined analogue to Subsection 8.3.2. Same as for calibration with test fields (Subsection 8.2.3), the hypothesis tests are carried out as Welch's t-test [Welch, 1947] as the variances and redundancies of $\hat{\boldsymbol{X}}_{I,c}$ and $\hat{\boldsymbol{X}}_{I,r}$ are not equal. The significance level is set to the common value of $\alpha = 5\%$. Non-significant deviations are of particular interest as they can be an indicator for reliable camera calibration with the proposed method. Tests on the deviations of standard deviations are not carried out, as it is desirable that the proposed method has better precision than the reference calibration and therefore a test on equality does not seem reasonable.

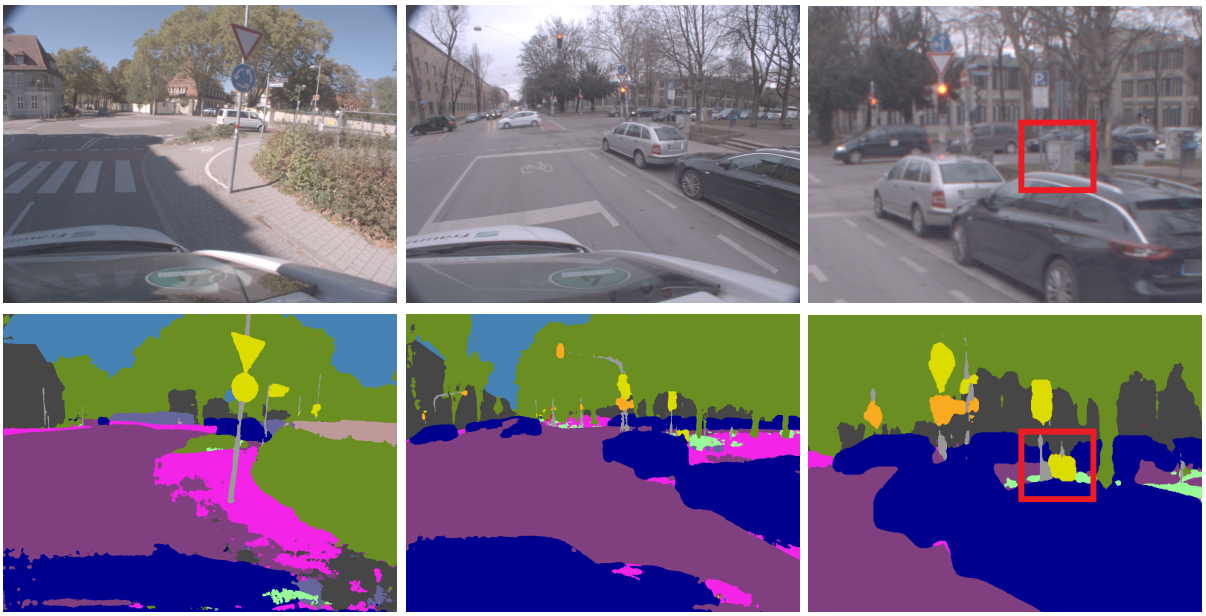
The hypothesis tests (Table 8.6) show for the *Ettlingen sequence* that there is the same number of non-significant deviations for k and p as for f and c. Non-significant deviations occur mostly for *1 C* and *1 CRT*. The hypothesis tests show that there are also for the *Munich sequence* more non-significant deviations for k and p than for c and f. Likewise, most non-significant deviations can be observed for the experimental case *1 C* and second-most for *1 CRT*. Overall, more non-significant deviations are obtained for the *Munich sequence* than for the *Ettlingen sequence*. In tendency, experimental cases with circular traffic signs provide more non-significant deviations than cases with triangular or rectangular signs, even though the plotted deviations for circular traffic signs (Figure 8.11) are not remarkably smaller compared to the other traffic sign shapes.

Table 8.6: Results of significance tests on the estimated interior orientation parameter values from calibration with traffic signs in comparison to the reference calibration for both test sequences for all experimental cases (C: circular traffic sign shape, R: rectangular shape, T: triangular shape; 1,2,3: different semantic segmentation and depth estimation methods). Non-significant deviation for value 0, significant deviation otherwise.

	Ettlingen sequence								Munich sequence							
Experimental case	f_x	f_y	c_x	c_y	k_1	k_2	p_1	p_2	f_x	f_y	c_x	c_y	k_1	k_2	p_1	p_2
1 C	0	0	0			0				0			0	0	0	
1 R																
1 T																
1 CR																
1 CT													0	0		
1 RT						0					0					0
1 CRT			0		0	0		0			0				0	0
2 CRT			0												0	
3 CRT															0	

8.3.4 Discussion

The experiments have shown that camera calibration with traffic signs has several major aspects that may influence the quality of calibration: First, the traffic sign contour needs to be determined precisely. Therefore, semantic segmentation with good boundary quality is necessary to detect traffic signs (Figure 8.12a). Small, i.e. far away, or partly occluded traffic signs can give false negative detections or have bad boundaries. Multiple signs next to each other might be identified as a single one and so the boundary may be wrong as well (Figure 8.12b). Both aspects are prob-

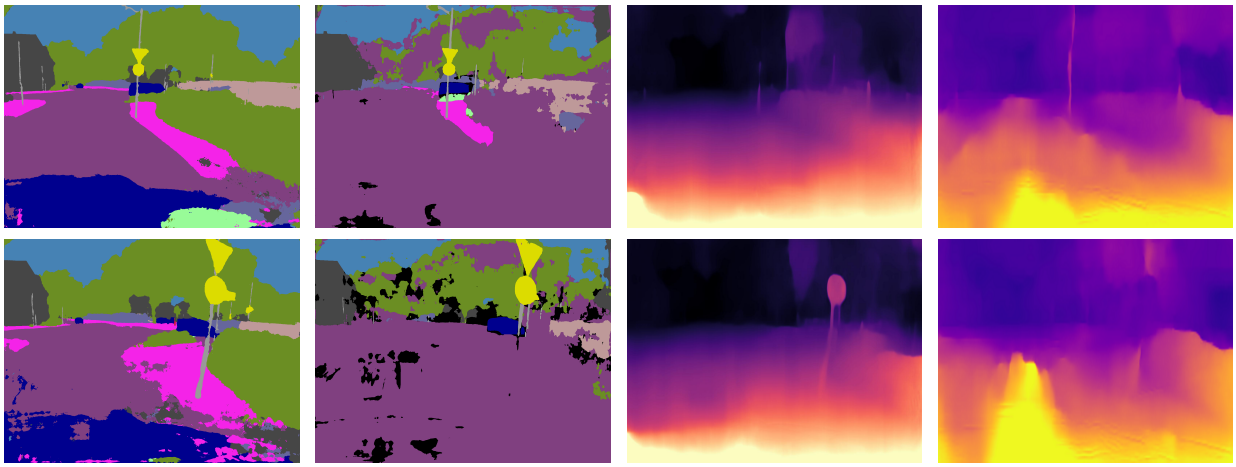


(a) (b) (c)

Figure 8.12: Examples for semantic images for calibration with traffic signs. a) *Ettlingen sequence*, two traffic sign segments with precise boundaries in the foreground, b) *Munich sequence*, multiple traffic signs contained in one segment with a bad boundary of the signs, c) *Munich sequence*, a false positive traffic sign detection (red rectangle) at a gray electric box at the edge of a sidewalk (cropped from the image in b).

lematic, as missed signs can't provide reference points and wrong boundaries can lead to wrong reference point coordinates. In contrast, also false positive detections are undesired, as they will provide invalid reference points (Figure 8.12c). In general, there might be remarkable differences in the semantic images, thus careful selection of an appropriate method for semantic segmentation is crucial (Figure 8.13a,b; note esp. the sidewalk in pink). After semantic segmentation providing initial boundaries, robust image processing is necessary to determine the right shape of a detected traffic sign (triangle, rectangle, circle) and then to precisely determine its boundary. Especially problematic cases like mentioned above impose the risk of false or imprecise boundaries, wherefore outlier removal at several steps has been proven in several experiments to be useful (e.g. based on the minimal area of the traffic sign segment, minimal overlap between traffic sign boundaries and semantic segments, maximal residual threshold during RANSAC, maximal color deviations of a traffic sign between RGB image and governmental color regulations). Thereby, precise boundaries (examples see Figure 8.14a,b,d top row) can be obtained even for bad initial boundaries from semantic segmentation.

Second, traffic signs have to be covered adequately by the deep models for depth estimation to obtain accurate distance values between camera and signs, which is fulfilled for the proposed method if at least a part of a traffic sign is visible in the depth image (Figure 8.13c). As observed by visual analysis, the tested methods for depth estimation tend to miss small foreground objects in contrast to large objects like trees or the road surface. Therefore, in some depth images, traffic signs might be even missing completely (Figure 8.13d) and so the determined distance will be wrong.



(a) (b) (c) (d)

Figure 8.13: Differences between two methods for semantic segmentation and depth estimation for the *Ettlingen sequence*. a) Semantic segmentation with Deeplabv3+ [Chen et al., 2018b], b) panoptic segmentation with EfficientPS [Mohan & Valada, 2021], c) depth estimation with monodepth2 [Godard et al., 2019], d) depth estimation with struct2depth [Casser et al., 2019].

Third, experiments have shown that the used boundary detection method tends to produce cluttered outer edges of the traffic sign boundaries because of other objects in the proximity of traffic signs (Figure 8.14 bottom row, traffic sign boundaries in white color). Therefore, it became necessary to use the non-cluttered inner edge for boundary extraction, which is displaced from the desired actual boundary of the traffic sign and so an empirically determined offset is applied for correction. Forth, the metric size of the traffic sign needs to be known for correct scaling of the reference point coordinates in the object space. If, as for example in Germany, the size of a sign depends on the prevalent speed limit, the size needs to be determined individually for each street. For the test sequences, it was most feasible to determine the size manually from official

regulations, but for large-scale use, this step needs to be automated, for example by querying speed limit information from street maps. Especially the estimated focal length strongly depends on depth estimation and the correct metric size of the traffic signs. Revealing systematic errors in the size with the available data is hardly possible. Either additional data, e.g. positions from GPS, could resolve this problem or large image sequences could alleviate the effect of such errors. Fifth, it has to be acknowledged that many German traffic sign shapes (e.g. stop sign) are not supported by the proposed method yet, in addition to traffic signs from other countries. Achieving a desirably large set of reference points would therefore require high alignment effort to the shapes of traffic signs in the desired area of application. Sixth, visual analysis of both test image sequences has shown that reference points have been obtained mostly for images acquired during the approach to an intersection, which can be an explanation for the higher number of reference points in the urban *Munich sequence*. In contrast, there are many images taken driving along straight street sections, where no traffic signs have been detected. Such images could be excluded easily from processing (e.g. by a known vehicle position) to save computational power and to reduce the risk of getting invalid reference points.

(a) (b) (c) (d)

Figure 8.14: Examples for RGB images with detected traffic signs and for boundary images from both test sequences. Same images as in Figures 8.12 and 8.13. Coarse boundary images (bottom row) obtained by CASENet [Yu et al., 2017]. The inner edge of the traffic sign boundary often represents the shape more precisely, while the outer edge may be cluttered because of nearby objects. a) Remote traffic signs in the *Ettlingen sequence* (cropped), b) close traffic signs in the *Ettlingen sequence*, c) remote traffic signs in the *Munich sequence* (cropped), d) close traffic signs in the *Munich sequence* (cropped).

8.4 Camera calibration by semantic structure-from-motion

Also for camera calibration by semantic structure-from-motion, the evaluation comprises statistical measures, deviation plots and hypothesis tests. All shown values are averages from 10 experimental runs. Same as for evaluation of calibration with traffic signs (Section 8.3), analysis is carried out in comparison among different experimental cases of the proposed method and with regard to the reference calibration.

8.4.1 Statistics

For both the *Ettlingen* and the *Munich sequence*, the same statistical measures as for camera calibration with traffic signs (Subsection 8.3.1) are presented. Additionally, the number of raw matches (#matches) obtained by semantic matching between the extracted features is shown.

Table 8.7: Statistical measures for calibration by semantic structure-from-motion for the *Ettlingen sequence* for all experimental cases. $\bar{r}$ represents the mean residuals of the image points in x- and y-direction and σ_r represents the standard deviation of the residuals in x- and y-direction. As some experimental cases cover only parts of the workflow, some cells are empty.

Experimental case	#features	#matches	#image points	#object points	$\bar{r}_x$ [px]	$\bar{r}_y$ [px]	σ_{r_x} [px]	σ_{r_y} [px]
MT1	3,481,185	4,686,271	4,394,528	123,925	$-3.60 \cdot 10^{-09}$	$-6.39 \cdot 10^{-09}$	$5.88 \cdot 10^{-01}$	$5.71 \cdot 10^{-01}$
MT2 + SFE	3,307,096	4,256,391	4,080,522	120,778	$-6.16 \cdot 10^{-09}$	$-6.48 \cdot 10^{-09}$	$5.82 \cdot 10^{-01}$	$5.63 \cdot 10^{-01}$
MT3 + SFE	823,901	1,330,282	1,230,210	48,992	$-1.75 \cdot 10^{-09}$	$-2.27 \cdot 10^{-09}$	$6.06 \cdot 10^{-01}$	$5.75 \cdot 10^{-01}$
MT4 + SFE	3,371,210	4,401,435	4,207,305	123,766	$-5.27 \cdot 10^{-09}$	$-7.22 \cdot 10^{-09}$	$5.83 \cdot 10^{-01}$	$5.68 \cdot 10^{-01}$
MT5 + SFE	3,386,609	4,479,131	4,280,607	124,160	$-4.28 \cdot 10^{-09}$	$-5.87 \cdot 10^{-09}$	$5.88 \cdot 10^{-01}$	$5.71 \cdot 10^{-01}$
MT6 + SFE	3,137,304	3,894,407	3,717,773	113,814	$-3.54 \cdot 10^{-09}$	$-3.82 \cdot 10^{-09}$	$5.78 \cdot 10^{-01}$	$5.62 \cdot 10^{-01}$
MT7 + SFE	2,943,201	4,015,788	3,881,496	89,134	$-8.70 \cdot 10^{-09}$	$-8.96 \cdot 10^{-10}$	$5.86 \cdot 10^{-01}$	$5.87 \cdot 10^{-01}$
MT8 + SFE	3,363,605	4,431,139	4,239,245	123,211	$-4.33 \cdot 10^{-09}$	$-5.08 \cdot 10^{-09}$	$5.86 \cdot 10^{-01}$	$5.68 \cdot 10^{-01}$
MT9 + SFE	3,304,884	4,252,997	4,077,714	120,917	$-6.23 \cdot 10^{-09}$	$-6.60 \cdot 10^{-09}$	$5.80 \cdot 10^{-01}$	$5.61 \cdot 10^{-01}$
MT10 + SFE	761,830	1,063,757	996,624	47,840	$1.42 \cdot 10^{-09}$	$-7.35 \cdot 10^{-10}$	$5.77 \cdot 10^{-01}$	$5.25 \cdot 10^{-01}$
MT1 + SM		4,575,547	4,241,815	120,594	$-1.60 \cdot 10^{-09}$	$-1.86 \cdot 10^{-09}$	$5.86 \cdot 10^{-01}$	$5.78 \cdot 10^{-01}$
MT2 + SFE + SM		4,158,305	3,954,721	117,004	$-2.27 \cdot 10^{-09}$	$-3.61 \cdot 10^{-09}$	$5.80 \cdot 10^{-01}$	$5.68 \cdot 10^{-01}$
MT3 + SFE + SM		1,276,044	1,158,731	47,686	$-2.97 \cdot 10^{-09}$	$-4.99 \cdot 10^{-10}$	$6.11 \cdot 10^{-01}$	$5.91 \cdot 10^{-01}$
MT4 + SFE + SM		4,300,194	4,073,125	120,150	$-2.92 \cdot 10^{-09}$	$-3.76 \cdot 10^{-09}$	$5.85 \cdot 10^{-01}$	$5.76 \cdot 10^{-01}$
MT5 + SFE + SM		4,367,409	4,133,949	120,787	$-2.30 \cdot 10^{-09}$	$-2.93 \cdot 10^{-09}$	$5.88 \cdot 10^{-01}$	$5.79 \cdot 10^{-01}$
MT6 + SFE + SM		3,866,257	3,663,835	110,661	$-4.31 \cdot 10^{-10}$	$-8.18 \cdot 10^{-11}$	$5.79 \cdot 10^{-01}$	$5.70 \cdot 10^{-01}$
MT7 + SFE + SM		3,923,154	3,768,316	89,889	$-2.62 \cdot 10^{-09}$	$-1.78 \cdot 10^{-09}$	$5.87 \cdot 10^{-01}$	$5.91 \cdot 10^{-01}$
MT8 + SFE + SM		4,326,927	4,102,854	119,282	$-4.33 \cdot 10^{-10}$	$-6.05 \cdot 10^{-10}$	$5.88 \cdot 10^{-01}$	$5.76 \cdot 10^{-01}$
MT9 + SFE + SM		4,155,999	3,952,470	116,820	$-1.65 \cdot 10^{-09}$	$-2.32 \cdot 10^{-09}$	$5.78 \cdot 10^{-01}$	$5.68 \cdot 10^{-01}$
MT10 + SFE + SM		1,017,090	937,896	45,423	$1.74 \cdot 10^{-08}$	$9.97 \cdot 10^{-09}$	$5.78 \cdot 10^{-01}$	$5.39 \cdot 10^{-01}$
MT1 + VMM					$-3.78 \cdot 10^{-09}$	$-7.65 \cdot 10^{-09}$	$5.88 \cdot 10^{-01}$	$5.71 \cdot 10^{-01}$
MT2 + SFE + VMM					$-6.34 \cdot 10^{-09}$	$-7.48 \cdot 10^{-09}$	$5.82 \cdot 10^{-01}$	$5.63 \cdot 10^{-01}$
MT3 + SFE +VMM					$-2.20 \cdot 10^{-09}$	$-4.10 \cdot 10^{-09}$	$6.06 \cdot 10^{-01}$	$5.75 \cdot 10^{-01}$
MT4 + SFE +VMM					$-5.41 \cdot 10^{-09}$	$-8.60 \cdot 10^{-09}$	$5.83 \cdot 10^{-01}$	$5.68 \cdot 10^{-01}$
MT5 + SFE +VMM					$-4.47 \cdot 10^{-09}$	$-6.86 \cdot 10^{-09}$	$5.88 \cdot 10^{-01}$	$5.71 \cdot 10^{-01}$
MT6 + SFE +VMM					$-3.61 \cdot 10^{-09}$	$-4.14 \cdot 10^{-09}$	$5.78 \cdot 10^{-01}$	$5.62 \cdot 10^{-01}$
MT7 + SFE +VMM					$-8.91 \cdot 10^{-09}$	$-1.80 \cdot 10^{-09}$	$5.86 \cdot 10^{-01}$	$5.87 \cdot 10^{-01}$
MT8 + SFE +VMM					$-4.46 \cdot 10^{-09}$	$-6.06 \cdot 10^{-09}$	$5.86 \cdot 10^{-01}$	$5.68 \cdot 10^{-01}$
MT9 + SFE +VMM					$-6.34 \cdot 10^{-09}$	$-7.99 \cdot 10^{-09}$	$5.80 \cdot 10^{-01}$	$5.61 \cdot 10^{-01}$
MT10 + SFE +VMM					$1.42 \cdot 10^{-09}$	$-1.24 \cdot 10^{-09}$	$5.77 \cdot 10^{-01}$	$5.25 \cdot 10^{-01}$
MT1 + SM + VMM					$-1.96 \cdot 10^{-09}$	$-2.21 \cdot 10^{-09}$	$5.86 \cdot 10^{-01}$	$5.78 \cdot 10^{-01}$
MT2 + SFE + SM + VMM					$-2.18 \cdot 10^{-09}$	$-4.20 \cdot 10^{-09}$	$5.80 \cdot 10^{-01}$	$5.68 \cdot 10^{-01}$
MT3 + SFE + SM + VMM					$-2.91 \cdot 10^{-09}$	$-2.59 \cdot 10^{-09}$	$6.11 \cdot 10^{-01}$	$5.91 \cdot 10^{-01}$
MT4 + SFE + SM + VMM					$-3.06 \cdot 10^{-09}$	$-4.43 \cdot 10^{-09}$	$5.85 \cdot 10^{-01}$	$5.76 \cdot 10^{-01}$
MT5 + SFE + SM + VMM					$-2.25 \cdot 10^{-09}$	$-3.21 \cdot 10^{-09}$	$5.88 \cdot 10^{-01}$	$5.79 \cdot 10^{-01}$
MT6 + SFE + SM + VMM					$-2.95 \cdot 10^{-10}$	$-4.08 \cdot 10^{-10}$	$5.79 \cdot 10^{-01}$	$5.70 \cdot 10^{-01}$
MT7 + SFE + SM + VMM					$-2.70 \cdot 10^{-09}$	$-2.33 \cdot 10^{-09}$	$5.87 \cdot 10^{-01}$	$5.91 \cdot 10^{-01}$
MT8 + SFE + SM + VMM					$-2.89 \cdot 10^{-10}$	$-6.92 \cdot 10^{-10}$	$5.88 \cdot 10^{-01}$	$5.76 \cdot 10^{-01}$
MT9 + SFE + SM + VMM					$-1.54 \cdot 10^{-09}$	$-3.19 \cdot 10^{-09}$	$5.78 \cdot 10^{-01}$	$5.68 \cdot 10^{-01}$
MT10 + SFE + SM + VMM					$1.65 \cdot 10^{-08}$	$1.26 \cdot 10^{-09}$	$5.78 \cdot 10^{-01}$	$5.39 \cdot 10^{-01}$

The number of image points (#image points) of the reference points represents the number of geometrically verified matches, whereby verification is done during 3d reconstruction by robust estimation of the homography, essential matrix or fundamental matrix between image pairs and serves to determine sets of feature points leading to valid estimations [Schönberger & Frahm, 2016]. The number of objects points (#object points) of the reference points corresponds with the number of 3d points of the sparse point cloud obtained by 3d reconstruction. The statistical measures are shown for four groups of experiments, all with semantic feature extraction (SFE), but 1) without semantic matching (SM) and without vehicle motion model (VMM), 2) with SM only, 3) with VMM only and 4) with both SM and VMM. Each group covers 10 experimental cases, one for each mask type. Obviously, for experimental cases with VMM only the residual measures are shown as these cases cover only the later part of the workflow (cf. Subsection 7.2.3).

About the number of extracted features, feature matches, image and object points obtained for the *Ettlingen sequence*, the following remarkable observations can be made (Table 8.7). First,

Table 8.8: Statistical measures for calibration by semantic structure-from-motion for the *Munich sequence* for all experimental cases. $\bar{r}$ represents the mean residuals of the image points in x- and y-direction and σ_r represents the standard deviation of the residuals in x- and y-direction. As some experimental cases cover only parts of the workflow, some cells are empty.

Experimental case	#features	#matches	#image points	#object points	$\bar{r}_x$ [px]	$\bar{r}_y$ [px]	σ_{r_x} [px]	σ_{r_y} [px]
MT1	4,083,494	11,303,684	10,782,210	72,697	$9.39 \cdot 10^{-08}$	$-9.68 \cdot 10^{-08}$	$6.47 \cdot 10^{-01}$	$6.68 \cdot 10^{-01}$
MT2 + SFE	3,532,771	8,653,334	8,396,234	55,383	$1.42 \cdot 10^{-03}$	$-6.40 \cdot 10^{-04}$	$6.34 \cdot 10^{-01}$	$6.70 \cdot 10^{-01}$
MT3 + SFE	1,252,205	6,152,047	5,906,410	54,617	$3.87 \cdot 10^{-08}$	$-2.57 \cdot 10^{-07}$	$6.55 \cdot 10^{-01}$	$6.39 \cdot 10^{-01}$
MT4 + SFE	3,932,046	10,115,536	9,779,667	72,132	$5.10 \cdot 10^{-08}$	$1.95 \cdot 10^{-08}$	$6.56 \cdot 10^{-01}$	$6.75 \cdot 10^{-01}$
MT5 + SFE	3,944,494	10,344,861	10,006,369	71,962	$6.99 \cdot 10^{-08}$	$2.31 \cdot 10^{-08}$	$6.53 \cdot 10^{-01}$	$6.71 \cdot 10^{-01}$
MT6 + SFE	3,336,142	6,866,702	6,655,381	49,053	$1.37 \cdot 10^{-07}$	$-1.60 \cdot 10^{-07}$	$6.66 \cdot 10^{-01}$	$7.12 \cdot 10^{-01}$
MT7 + SFE	3,852,804	9,930,642	9,624,462	64,293	$2.18 \cdot 10^{-08}$	$1.57 \cdot 10^{-08}$	$6.46 \cdot 10^{-01}$	$6.76 \cdot 10^{-01}$
MT8 + SFE	3,904,666	10,094,971	9,759,739	71,468	$1.68 \cdot 10^{-07}$	$3.16 \cdot 10^{-08}$	$6.57 \cdot 10^{-01}$	$6.74 \cdot 10^{-01}$
MT9 + SFE	3,518,077	8,624,113	8,369,520	55,262	$3.51 \cdot 10^{-08}$	$2.26 \cdot 10^{-08}$	$6.34 \cdot 10^{-01}$	$6.71 \cdot 10^{-01}$
MT10 + SFE	801,547	4,173,223	4,024,506	36,580	$1.10 \cdot 10^{-07}$	$4.10 \cdot 10^{-08}$	$6.19 \cdot 10^{-01}$	$6.01 \cdot 10^{-01}$
MT1 + SM		11,206,896	10,628,034	78,600	$-2.05 \cdot 10^{-07}$	$8.08 \cdot 10^{-08}$	$6.59 \cdot 10^{-01}$	$6.82 \cdot 10^{-01}$
MT2 + SFE + SM		8,536,379	8,244,000	59,246	$1.85 \cdot 10^{-07}$	$9.62 \cdot 10^{-08}$	$6.37 \cdot 10^{-01}$	$6.79 \cdot 10^{-01}$
MT3 + SFE + SM		6,163,664	5,882,881	57,416	$3.57 \cdot 10^{-08}$	$5.87 \cdot 10^{-09}$	$6.65 \cdot 10^{-01}$	$6.54 \cdot 10^{-01}$
MT4 + SFE + SM		10,120,195	9,724,269	78,850	$4.99 \cdot 10^{-08}$	$-9.05 \cdot 10^{-10}$	$6.63 \cdot 10^{-01}$	$6.86 \cdot 10^{-01}$
MT5 + SFE + SM		10,322,060	9,922,048	78,779	$1.31 \cdot 10^{-07}$	$-3.88 \cdot 10^{-07}$	$6.59 \cdot 10^{-01}$	$6.81 \cdot 10^{-01}$
MT6 + SFE + SM		6,937,273	6,682,712	50,601	$4.02 \cdot 10^{-08}$	$-6.78 \cdot 10^{-08}$	$6.75 \cdot 10^{-01}$	$7.23 \cdot 10^{-01}$
MT7 + SFE + SM		9,925,907	9,565,555	71,939	$1.11 \cdot 10^{-07}$	$7.49 \cdot 10^{-08}$	$6.58 \cdot 10^{-01}$	$6.89 \cdot 10^{-01}$
MT8 + SFE + SM		10,104,901	9,711,579	76,648	$-1.70 \cdot 10^{-07}$	$-2.06 \cdot 10^{-08}$	$6.67 \cdot 10^{-01}$	$6.87 \cdot 10^{-01}$
MT9 + SFE + SM		8,502,832	8,215,449	58,385	$1.84 \cdot 10^{-07}$	$6.19 \cdot 10^{-08}$	$6.34 \cdot 10^{-01}$	$6.73 \cdot 10^{-01}$
MT10 + SFE + SM		4,132,805	3,971,187	38,071	$1.44 \cdot 10^{-07}$	$3.43 \cdot 10^{-09}$	$6.19 \cdot 10^{-01}$	$6.10 \cdot 10^{-01}$
MT1 + VMM					$6.53 \cdot 10^{-08}$	$2.19 \cdot 10^{-08}$	$6.47 \cdot 10^{-01}$	$6.68 \cdot 10^{-01}$
MT2 + SFE + VMM					$1.81 \cdot 10^{-03}$	$-4.30 \cdot 10^{-04}$	$6.34 \cdot 10^{-01}$	$6.70 \cdot 10^{-01}$
MT3 + SFE +VMM					$1.83 \cdot 10^{-07}$	$8.96 \cdot 10^{-08}$	$6.56 \cdot 10^{-01}$	$6.40 \cdot 10^{-01}$
MT4 + SFE +VMM					$8.73 \cdot 10^{-08}$	$2.36 \cdot 10^{-08}$	$6.56 \cdot 10^{-01}$	$6.75 \cdot 10^{-01}$
MT5 + SFE +VMM					$5.11 \cdot 10^{-08}$	$2.00 \cdot 10^{-08}$	$6.52 \cdot 10^{-01}$	$6.71 \cdot 10^{-01}$
MT6 + SFE +VMM					$4.11 \cdot 10^{-08}$	$-1.78 \cdot 10^{-09}$	$6.64 \cdot 10^{-01}$	$7.10 \cdot 10^{-01}$
MT7 + SFE +VMM					$5.07 \cdot 10^{-08}$	$2.28 \cdot 10^{-08}$	$6.46 \cdot 10^{-01}$	$6.76 \cdot 10^{-01}$
MT8 + SFE +VMM					$6.95 \cdot 10^{-08}$	$3.54 \cdot 10^{-08}$	$6.57 \cdot 10^{-01}$	$6.74 \cdot 10^{-01}$
MT9 + SFE +VMM					$8.81 \cdot 10^{-08}$	$1.77 \cdot 10^{-08}$	$6.34 \cdot 10^{-01}$	$6.71 \cdot 10^{-01}$
MT10 + SFE +VMM					$2.25 \cdot 10^{-08}$	$7.53 \cdot 10^{-09}$	$6.19 \cdot 10^{-01}$	$6.01 \cdot 10^{-01}$
MT1 + SM + VMM					$3.82 \cdot 10^{-08}$	$1.61 \cdot 10^{-08}$	$6.59 \cdot 10^{-01}$	$6.82 \cdot 10^{-01}$
MT2 + SFE + SM + VMM					$1.12 \cdot 10^{-07}$	$1.40 \cdot 10^{-08}$	$6.37 \cdot 10^{-01}$	$6.79 \cdot 10^{-01}$
MT3 + SFE + SM + VMM					$2.63 \cdot 10^{-08}$	$7.14 \cdot 10^{-09}$	$6.64 \cdot 10^{-01}$	$6.53 \cdot 10^{-01}$
MT4 + SFE + SM + VMM					$5.21 \cdot 10^{-08}$	$1.62 \cdot 10^{-08}$	$6.64 \cdot 10^{-01}$	$6.86 \cdot 10^{-01}$
MT5 + SFE + SM + VMM					$1.11 \cdot 10^{-06}$	$-5.40 \cdot 10^{-07}$	$6.59 \cdot 10^{-01}$	$6.81 \cdot 10^{-01}$
MT6 + SFE + SM + VMM					$8.74 \cdot 10^{-10}$	$-1.44 \cdot 10^{-09}$	$6.76 \cdot 10^{-01}$	$7.24 \cdot 10^{-01}$
MT7 + SFE + SM + VMM					$5.12 \cdot 10^{-08}$	$2.79 \cdot 10^{-08}$	$6.58 \cdot 10^{-01}$	$6.88 \cdot 10^{-01}$
MT8 + SFE + SM + VMM					$5.93 \cdot 10^{-08}$	$-1.25 \cdot 10^{-07}$	$6.68 \cdot 10^{-01}$	$6.88 \cdot 10^{-01}$
MT9 + SFE + SM + VMM					$5.04 \cdot 10^{-08}$	$2.85 \cdot 10^{-08}$	$6.33 \cdot 10^{-01}$	$6.73 \cdot 10^{-01}$
MT10 + SFE + SM + VMM					$1.85 \cdot 10^{-07}$	$-3.59 \cdot 10^{-08}$	$6.19 \cdot 10^{-01}$	$6.10 \cdot 10^{-01}$

experimental cases with mask type 1 provide the highest numbers of extracted features, feature matches and verified matches. This fits to the expectation, as mask type 1 defines the baseline calibration by using an empty mask and so not excluding any image area from feature extraction. Second, for mask types 3 and 10, these numbers are clearly lower than for the other mask types (less than 50%), independent whether SM is used or not. Third, comparing the groups SFE and SFE + SM with each other, the number of matches, image and object points is lower for the later one for most experimental cases, but the differences are smaller than between some mask types. Forth, mask type 5 leads to the highest number of object points in all four groups, whereby the differences to mask types 1, 4 and 8 are less than 1%. The mean residuals $\bar{r}_x$ and $\bar{r}_y$ share the same exponent for most experimental cases, but the absolute values of $\bar{r}_y$ are often larger than the absolute values of $\bar{r}_x$. For any experimental case, the largest difference between $\bar{r}_x$ and $\bar{r}_y$ is around one power of ten. Even though almost all mean residuals have a negative sign, the largest exponent is -8, so both measures are below the typical image measurement accuracy

of 0.1 px. Therefore, no indication is seen for the presence of a systematic mean bias in the residuals. In contrast to the mean residuals, σ_{r_x} is larger than σ_{r_y} for most experimental cases. But as both kinds of measures have the same exponent in every experimental case, no direction-dependent differences in the value ranges of the residuals are seen. Comparing the groups with and without SM, the experimental cases in the later one have absolute values of $\bar{r}_x$ and $\bar{r}_y$ that are approximately $2-3$ times larger. For the standard deviations of the residuals, the differences are less and no tendency with regard to the use of SM can be observed. Comparing groups with and without VMM, $\bar{r}_x$ and $\bar{r}_y$ tend to be larger for groups with VMM, but the differences are typically smaller than one power of ten. σ_{r_x} and σ_{r_y} are, within the shown decimal digits, exactly the same. Comparing the mask types in different groups with each other, mask types 6, 8 and 10 provide the lowest $\bar{r}_x$ in some groups, and mask types 6 and 10 provide the lowest $\bar{r}_y$. Mask types 9 and 10 provide the lowest σ_{r_x} in at least one group and mask type 10 provides the lowest σ_{r_y} in all four groups.

For the *Munich sequence* (Table 8.8), the numbers of extracted features, feature matches, image points and object points are often lowest for mask type 10. These measures are highest for mask type 1 for all groups, except for the number of object points, which is highest for mask type 4 for the group with SM only. In comparison between the groups without and with SM, the number of matches and image points is lower for the later one for almost all experimental cases. This observation is in accordance with the *Ettlingen sequence*. However, the number of object points is higher for the SM groups, which is in opposite to the *Ettlingen sequence*. For most experimental cases, the exponents of the mean residuals are -7 and -8 for the *Munich sequence*, while they range in total from -3 to -10. Hence, the mean residuals are larger than for the *Ettlingen sequence*, but still below the image measurement accuracy. In contrast to the other sequence, the absolute values of $\bar{r}_x$ are larger than those of $\bar{r}_y$ and for most cases σ_{r_y} is larger than σ_{r_x}. Comparing the groups without and with SM with each other, $\bar{r}$ is smaller when SM is used for some experimental cases and larger for others. Comparing groups with and without VMM, $|\bar{r}_x|$ and $|\bar{r}_y|$ are smaller for VMM for some cases and larger for others. σ_{r_x} and σ_{r_y} are, within the shown decimal digits, the same for almost all experimental cases. Comparing the mask types among different groups with each other, mask types 3, 7 and 10 provide the lowest $|\bar{r}_x|$ in some groups, and mask types 4, 6 and 7 the lowest $|\bar{r}_y|$. Mask types 10 provides the lowest σ_{r_x} and σ_{r_y} in all groups. Finally, it can be concluded that observations on the use of SM, VMM and the mask types are sometimes contrary between the two sequences, which can indicate that the performance of the proposed method depends on the type of scene.

8.4.2 Deviations of orientation parameters between the proposed and a reference calibration

The same kind of deviations $\boldsymbol{d}_{cr,\boldsymbol{X}}$ for the estimated interior orientation parameter values and $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ for the estimated standard deviations are plotted as for calibration with traffic signs (cf. Subsection 8.3.2). Note the scaling of the vertical axis in the deviation plots has been selected so that differences between the experimental cases become clearly visible. Therefore, it became necessary to truncate extraordinary large deviations, which is why the corresponding bars of such off-limit deviations are set to a light gray color.

For $\boldsymbol{d}_{cr,\boldsymbol{X}}$ (Figures 8.15, 8.16), the deviations are in tendency larger for the distortion parameters than for the focal length and the principal point coordinates. For the focal length and the principal point coordinates, most deviations are between ± 0.1, which means that the estimated values deviate from reference calibration less than 10%. The y-components f_y and c_y show in tendency larger deviations than the x-components f_x and c_x, respectively. This applies especially for the *Munich sequence*, where the estimated values for f_y deviate approx. 25% in the maximal

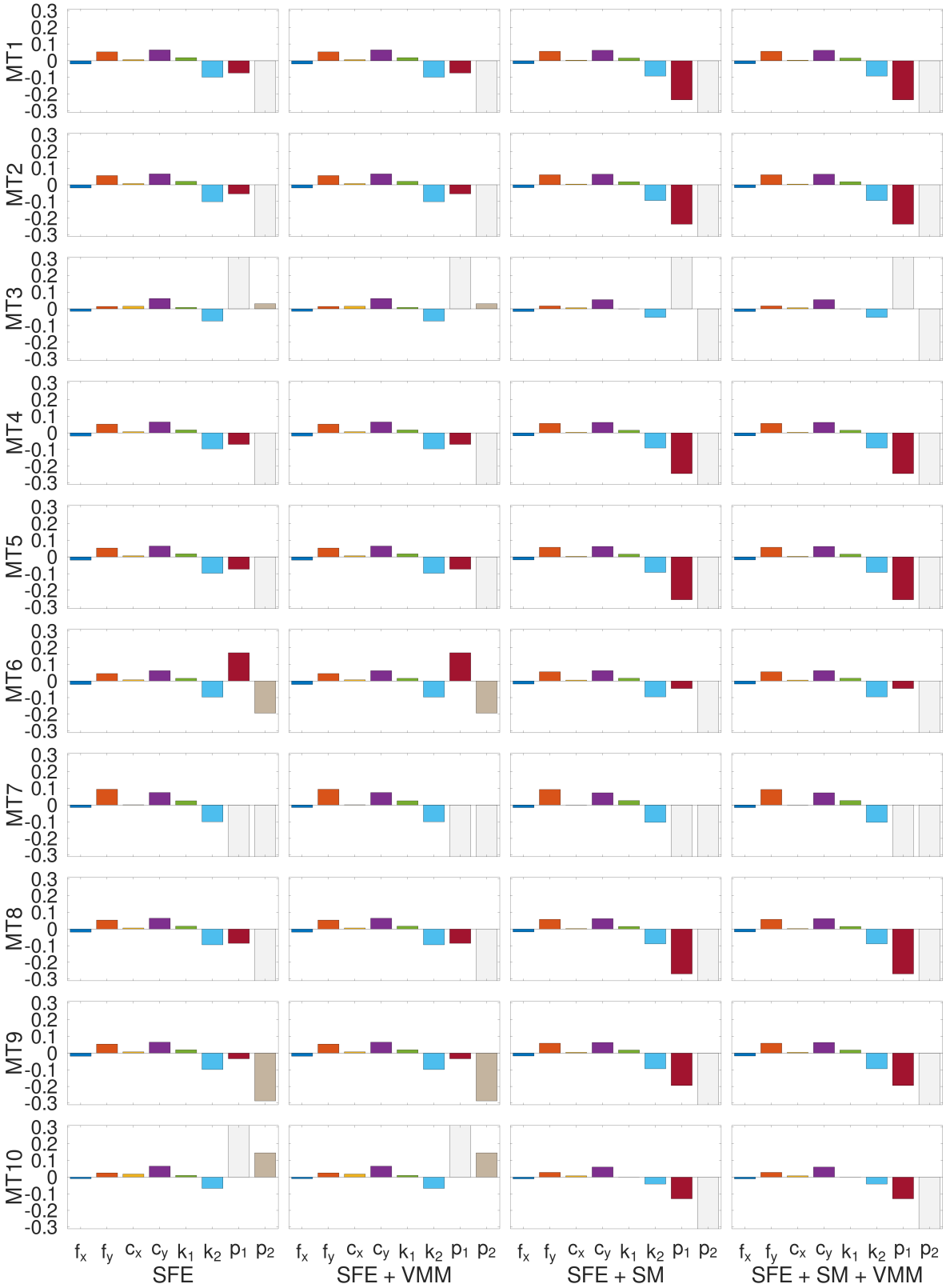


Figure 8.15: Deviations of the interior orientation parameter values obtained by calibration by semantic structure-from-motion from the reference calibration for the *Ettlingen sequence* for all experimental cases (MT: mask types, SFE: semantic feature extraction, SM: semantic matching, VMM: vehicle motion model). Vertical axis unit-free due to relative deviations, off-limit deviations in light gray.

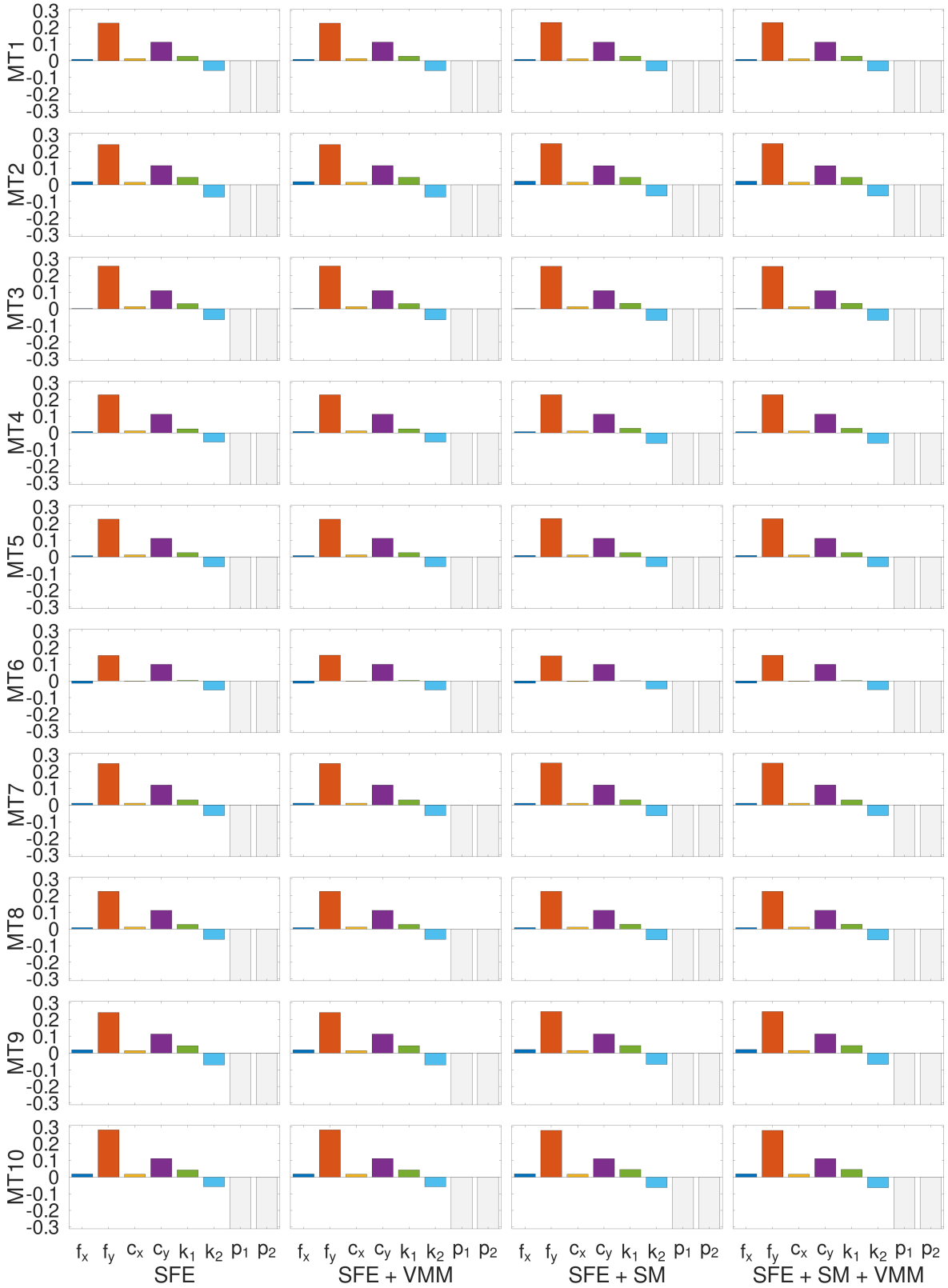


Figure 8.16: Deviations of the interior orientation parameter values obtained by calibration by semantic structure-from-motion from the reference calibration for the *Munich sequence* for all experimental cases (MT: mask types, SFE: semantic feature extraction, SM: semantic matching, VMM: vehicle motion model). Vertical axis unit-free due to relative deviations, off-limit deviations in light gray.

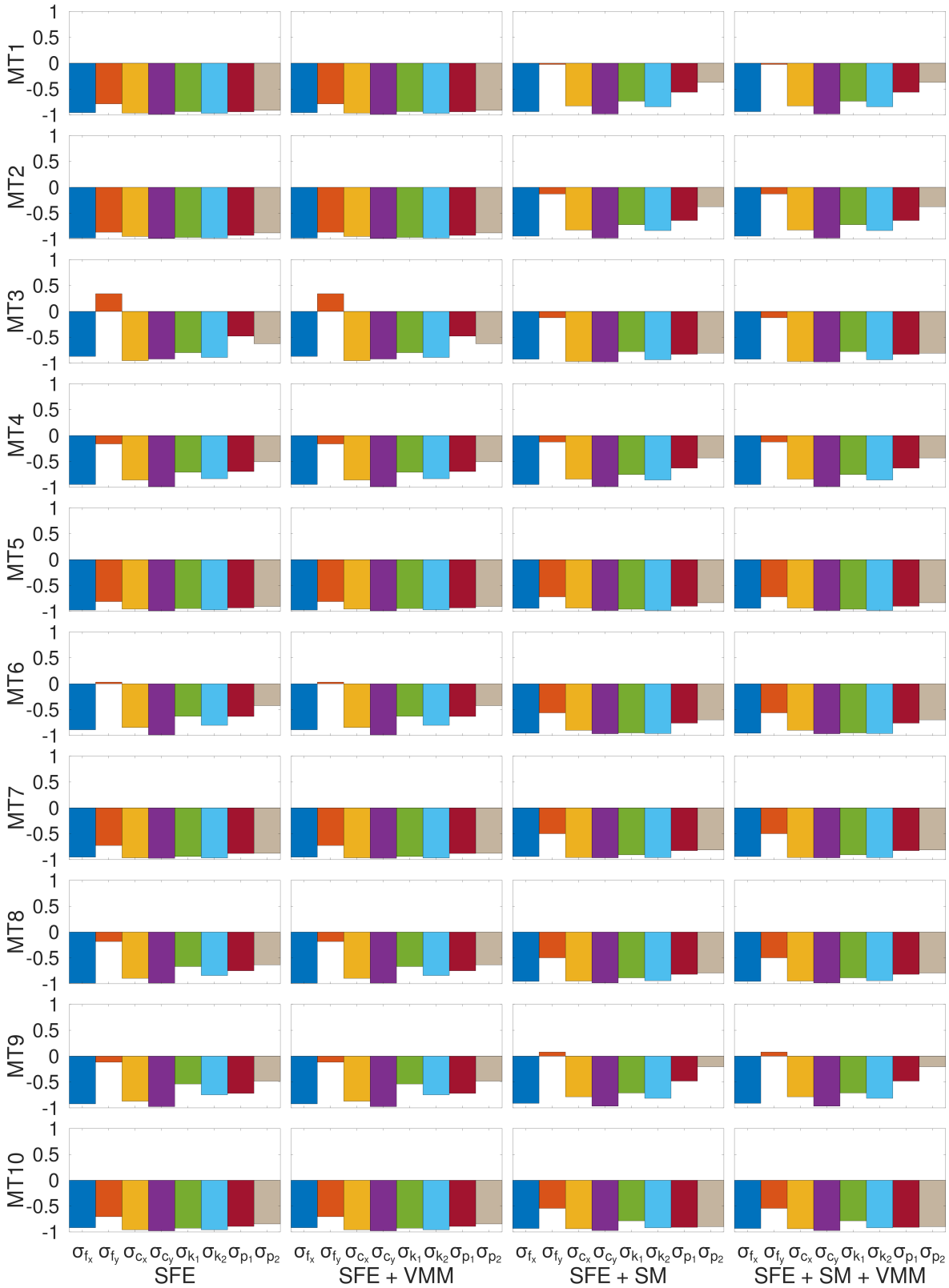


Figure 8.17: Deviations of the standard deviations of the interior orientation parameters obtained by calibration by semantic structure-from-motion from the reference calibration for the *Ettlingen sequence* for all experimental cases (MT: mask types, SFE: semantic feature extraction, SM: semantic matching, VMM: vehicle motion model). Vertical axis unit-free due to relative deviations, off-limit deviations in light gray.

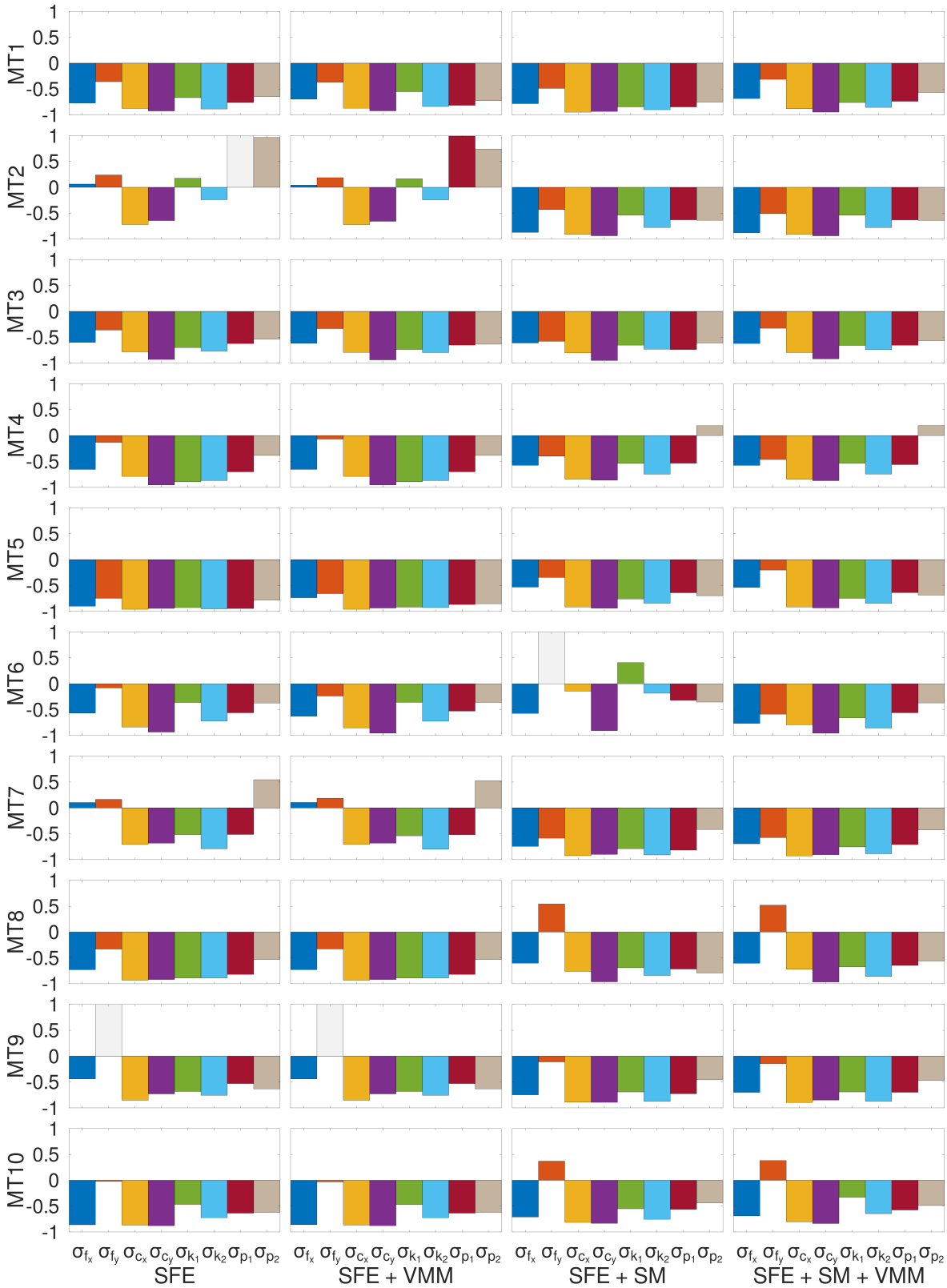


Figure 8.18: Deviations of the standard deviations of the interior orientation parameters obtained by calibration by semantic structure-from-motion from the reference calibration for the *Munich sequence* for all experimental cases (MT: mask types, SFE: semantic feature extraction, SM: semantic matching, VMM: vehicle motion model). Vertical axis unit-free due to relative deviations, off-limit deviations in light gray.

case from reference calibration. In average over all interior orientation parameters, the deviations are larger for the *Munich sequence* than for the *Ettlingen sequence*. Almost all deviations $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ (Figures 8.17, 8.18) are negative, hence, the precision of the proposed method can be considered as better than the precision of the reference calibration. While for some parameters (e.g. σ_{c_y}) the precision of the proposed method is clearly better ($\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ close to -1), it can be considered as slightly better or similar to reference calibration for other parameters (esp. f_y; $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ close to 0). *Ettlingen* shows in average larger negative deviations $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$ than *Munich*, hence a better precision.

Between groups with or without VMM, remarkable differences are observed only in a single experimental case (*Munich*, $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$, MT6 + SFE + SM). With regard to the use of SM, remarkable differences are observed only in a few experimental cases for $\boldsymbol{d}_{cr,\boldsymbol{X}}$: In some of them, experimental cases with SM have smaller deviations than those without SM (e.g. *Ettlingen*, MT10), in other cases they have larger ones (e.g. *Ettlingen*, MT2). Likewise for $\boldsymbol{d}_{cr,\sigma_{\boldsymbol{X}}}$, some experimental cases show larger negative deviations (better precision) with SM (e.g. *Munich*, MT2), while others show smaller negative deviations (worse precision) (e.g. *Ettlingen*, MT2 and MT9). Between most mask types, there are no remarkable differences visible. In a few experimental cases, some mask types (e.g. *Ettlingen*, MT7) show larger deviations $\boldsymbol{d}_{cr,\boldsymbol{X}}$ than the baseline with MT1, but the visual differences are small. At the same time, other mask types show, averaged over all interior orientation parameters, smaller deviations $\boldsymbol{d}_{cr,\boldsymbol{X}}$ than the baseline (e.g. *Ettlingen*, MT3 + SFE, MT6 + SFE + SM). These differences with regard to the baseline are often large for certain interior orientation parameters only (e.g. MT6 + SFE + SM, p_1), while they are not visible for other orientation parameters (same experimental case, f_x). Hence, these observations indicate a dependency of the effects of certain mask types on the specific orientation parameter.

8.4.3 Significance tests on deviations between estimated and reference orientation values

The hypothesis tests are carried out for the values of the interior orientation parameters estimated with calibration by semantic structure-from-motion in comparison to the reference calibration. The tests are defined analogue to the tests for calibration with traffic signs (cf. Subsection 8.3.3). Likewise, the key interest of the hypothesis tests is to determine which deviations between the proposed method and reference calibration are non-significant in a statistical meaning, which can be seen as indicator for reliable calibration. Due to the desired large negative values, there are no tests for the standard deviations (cf. Subsection 8.4.2).

The test results (Table 8.9) for the *Ettlingen sequence* show that 7.5% of all deviations are non-significant. The mask types with the highest rate of 12.5% non-significant deviations are 3 and 7. All other mask types have the same lower rate of 6.25%. For experimental cases with SM, less deviations are non-significant (5%) than for experimental cases without SM (10%). Comparing the groups with and without VMM, the rate of non-significant deviations is exactly the same (7.5%). The test results for the *Munich sequence* show that approx. 3% of all deviations are non-significant. The highest rate of approx. 15.6% non-significant deviations occur for mask type 6, followed by mask type 3 with 12.5%. For all other mask types, no deviation is statistically non-significant. Deviations for experimental cases with VMM are non-significant in approximately 3.1% of the cases, while it is 2.5% for cases without VMM. The same rates apply for experimental cases with and without SM, respectively. In contrast to the *Ettlingen sequence*, using semantic matching and the vehicle motion model leads to a slightly higher number of non-significant deviations for the *Munich sequence*.

Table 8.9: Results of significance tests on the estimated interior orientation parameter values for calibration by semantic structure-from-motion in comparison to the reference calibration for both test sequences for all experimental cases (MT: mask types, SFE: semantic feature extraction, SM: semantic matching, VMM: vehicle motion model). Non-significant deviation for value 0, significant deviation otherwise.

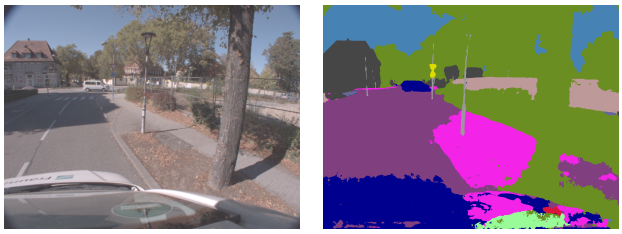
	Ettlingen sequence								Munich sequence							
Experimental case	f_x	f_y	c_x	c_y	k_1	k_2	p_1	p_2	f_x	f_y	c_x	c_y	k_1	k_2	p_1	p_2
MT1							0									
MT2 + SFE							0									
MT3 + SFE								0	0							
MT4 + SFE							0									
MT5 + SFE							0									
MT6 + SFE											0					
MT7 + SFE			0													
MT8 + SFE							0									
MT9 + SFE							0									
MT10 + SFE																
MT1 + SM																
MT2 + SFE + SM																
MT3 + SFE + SM					0				0							
MT4 + SFE + SM																
MT5 + SFE + SM																
MT6 + SFE + SM							0						0			
MT7 + SFE + SM			0													
MT8 + SFE + SM																
MT9 + SFE + SM																
MT10 + SFE + SM					0											
MT1 + VMM							0									
MT2 + SFE + VMM							0									
MT3 + SFE +VMM								0	0							
MT4 + SFE +VMM							0									
MT5 + SFE +VMM							0									
MT6 + SFE +VMM											0					
MT7 + SFE +VMM			0													
MT8 + SFE +VMM							0									
MT9 + SFE +VMM							0									
MT10 + SFE +VMM																
MT1 + SM + VMM																
MT2 + SFE + SM + VMM																
MT3 + SFE + SM + VMM					0				0							
MT4 + SFE + SM + VMM																
MT5 + SFE + SM + VMM																
MT6 + SFE + SM + VMM							0				0		0			
MT7 + SFE + SM + VMM			0													
MT8 + SFE + SM + VMM																
MT9 + SFE + SM + VMM																
MT10 + SFE + SM + VMM					0											

8.4.4 Discussion

Obviously, as first discussion aspect, the performance of semantic feature extraction or matching depends on the quality of semantic segmentation. Certain problems may arise that could lead to false semantic segmentation (Figure 8.19) and subsequently deteriorate the quality of camera

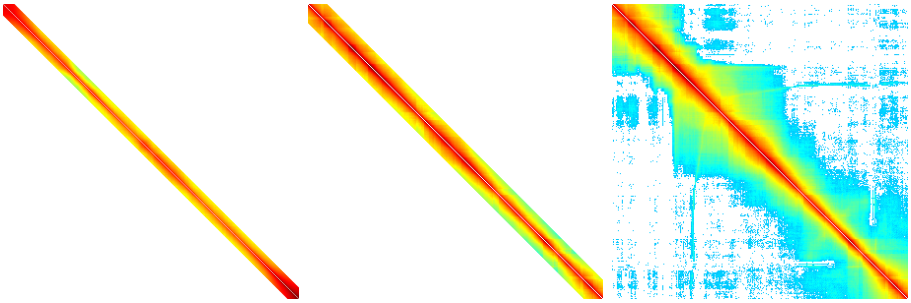
calibration. First, relevant characteristics of the images used for calibration may deviate from the images used for training the deep model for semantic segmentation. The images might be acquired with a different camera, under different scene illumination, from a different type of road scene, for instance. Fine tuning with calibration images could alleviate this problem, but would make the method dependent on the availability of labeled calibration images so that they can serve for supervised training. Second, rare or new objects like fast moving city scooters on side walks may not be represented by any trained semantic class at all. Third, the used semantic segmentation can't distinguish standing from moving objects and hence features on standing objects may be unnecessarily excluded. Forth, even unexpected classes might be relevant to exclude them from feature extraction to get a robust 3d reconstruction and so should be covered by semantic segmentation: As example, Kaneko et al. [2018] have found out in empirical studies that the class *sky* besides the class *car* is most relevant.

Second, in addition to the proposed "sequential matching", an "exhaustive matching" where the search for matches covers all possible combinations of images in the sequence has been tested without success (colored far-off-diagonals in Figure 8.20c). For some experimental cases, a remarkable number of wrong correspondences were established between images showing, for example, similar looking, but different streets (Figure 8.21d). In contrast, sequential matching that considers only certain intervals of images ordered by their acquisition time for matching does not suffer from this problem (white far-off-diagonals in Figure 8.20a,b) and so reliable matches are obtained in general (Figure 8.21a). Additionally, some images provide a lower number of matches compared to other images (yellow or green close to main diagonal in Figure 8.20). Such images will contribute with only a comparably small number of reference points and so they are more prone to have mismatches due to an insufficient number of other matches for strong geometric verification. Furthermore, the experiments have proven that inappropriate semantic masks or a missing fix-pixel mask for the ego-car can easily cause invalid matches in practice (Figure 8.21b,c).



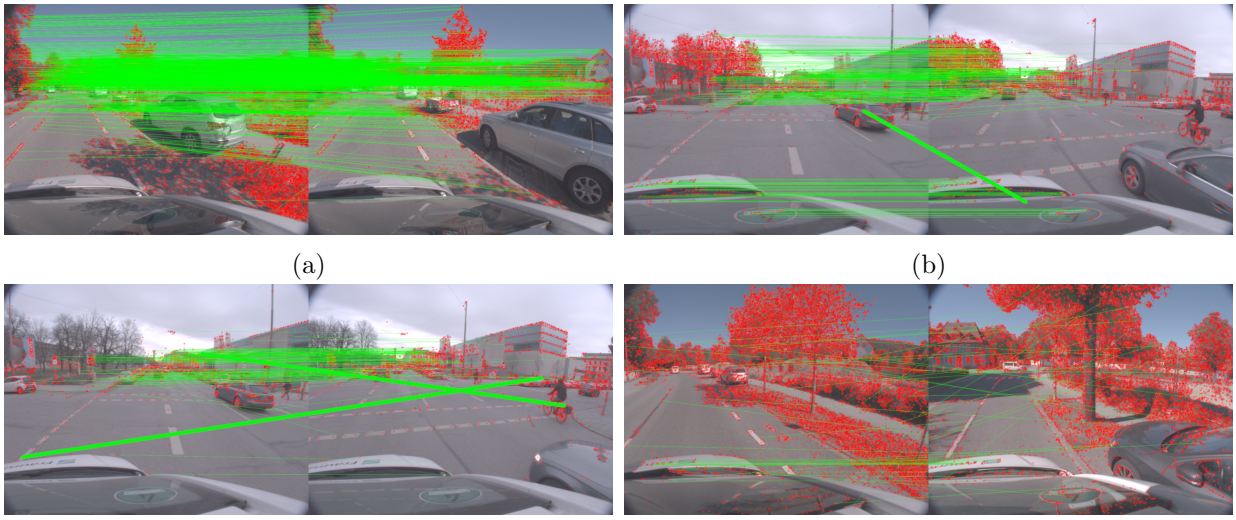
(a) (b)

Figure 8.19: Examples for wrong classes from semantic segmentation in the bottom right image quarter (mix of light green, red, dark blue, pink). a) RGB image, b) corresponding semantic image.



(a) (b) (c)

Figure 8.20: Match matrices for calibration by semantic structure-from-motion. Experimental case MT1, without SM and without VMM. One image per matrix row and column. Red color for high number of matches, blue for low. a) Sequential matching for the *Munich sequence* (504 images), b) sequential matching for the *Ettlingen sequence* (300 images), c) exhaustive matching for the *Ettlingen sequence*.



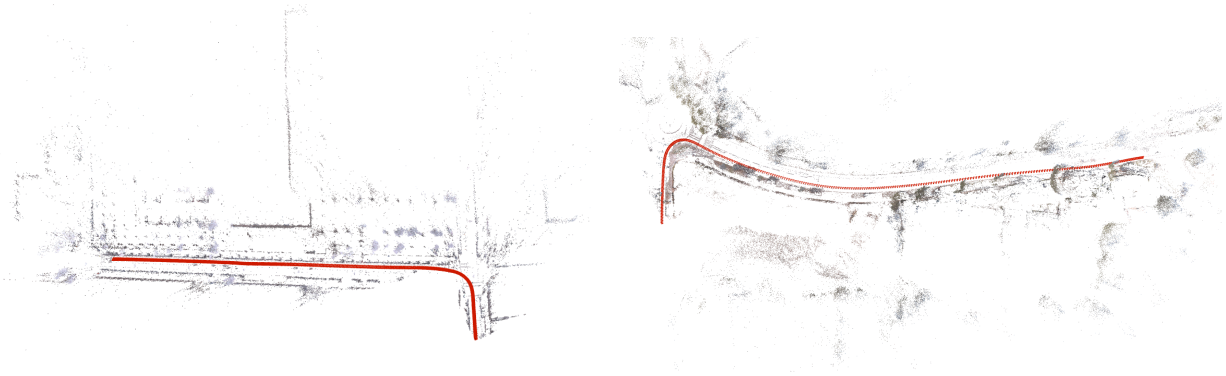
(a) (b)

(c) (d)

Figure 8.21: Various successfully matched image pairs with feature matches (green lines). a) Large number of valid matches between temporarily close images b) invalid match due to missing ego-car mask (thick green line), c) invalid match between different kinds of objects due to inappropriate semantic masks (thick green line), d) invalid association between temporarily distant images with no overlap for *exhaustive matching*.

Third, the large time amount of several minutes needed for creating the large 3d reconstructions (examples see Figure 8.22) and for bundle adjustment, even on high performance laboratory computers, may be a big challenge in cars with limited computational power. Real-time capability, i.e. that each image can be processed until the next one is acquired, is currently not achieved. Thus, selection of an appropriate sequence of the continuously recorded images for calibration becomes necessary. Besides the aspect of computational power, older images may not represent the current interior orientation anymore and hence this is another reason that should be considered in determining an appropriate calibration image sequence. Furthermore, bundle adjustment did not converge in some experimental runs; the residuals of such runs may easily be one hundred times as large as for converging runs, according to manual inspection. As far as analyzed, convergence failed if bad initial image pairs were selected or in the case of a bad order for registering the remaining images, which is both done randomly. Additionally, the origin of the object coordinate system in which the GPS positions used for the metric scale are provided needs to be defined close to the image positions so that the absolute values of the image positions and object coordinates of the reference points are small. If the origin was placed far off the images and reference points, convergence failed as well. Non-converging experimental runs were repeated.

Forth and last, the quality of camera calibration with the proposed method may suffer from difficulties affecting the acquisition geometry. First, movement of the camera needed for SfM is mainly along the optical axis due to the typical forward driving direction of a car, which could result in a poor 3d reconstruction, especially for objects that are far away from the camera. Second, the DOF coverage in the acquisition geometry is limited which could lead to poorly estimated parameters or large correlations between them as the proposed method relies on self-calibration [Luhmann et al., 2006]. For example, rigid camera mounting in the car allows no rotation wrt. the reference points, except by some degrees due to pitch and roll caused by the car suspension. Additionally, viewing angles on the reference points may be heavily restricted and are often similar in different images. For instance, top-down or bottom-up views on the reference points are not possible, obviously.



(a) (b)

Figure 8.22: 3d reconstructions have typically several ten thousand points that need to be handled by bundle adjustment to use them as reference for calibration. Top-down view. a) *Munich sequence*, b) *Ettlingen sequence*.

8.5 Comparative discussion of the proposed methods

Comparing the algorithms of both methods for camera self-calibration with each other, the following observations can be made. First, it can be expected that calibration by semantic structure-from-motion can be better generalized to unknown types of road scenes, which is important for the applicability of automotive camera calibration is mass-produced cars. Experiments have proven that image features, from which the reference points are obtained from, can be extracted from a large manifold of object types that can be typically found in many different types of scenes. In contrast, the appearance of traffic signs is a crucial factor for an algorithm to extract the reference point coordinates, thus the different appearance of traffic signs in different countries will be a limiting factor for generalization. Furthermore, calibration with traffic signs relies on a larger number of deep learning-based methods than calibration by semantic structure-from-motion, thus the risk that one of the models performs bad in a new type of road scene is higher. Additionally, experiments have shown that reference points from image features are distributed wider and more equal across the image area than reference points from traffic signs that are mostly in the upper image area. Thus, it can be expected that outliers resulting from a lack of generalization capability can be detected better by calibration by semantic structure-from-motion. Second, calibration by semantic structure-from-motion requires a lower number of additional information which may be difficult to obtain and may be an additional source for errors. While this method only needs the vehicle position that can be easily obtained from GPS, calibration with traffic signs needs the speed limit in order to determine the speed-dependent metric size of the traffic signs. Even though already two data sources are necessary to obtain the speed limit from GPS and a map, even a third data source is needed to obtain the metric size of traffic signs e.g. from official regulations for each country where this calibration method should be utilized. Third, the total processing time of the experiments has shown that the computational effort for calibration by semantic structure-from-motion is remarkably larger. Especially the incremental creation of the 3d point cloud requires more time than applying all deep models for calibration with traffic signs.

The results of both methods for camera self-calibration can be compared well with each other, as both have been tested with the same datasets and both have been evaluated against the same reference calibration. The results indicate major differences in the quality of both methods. First, as it can be expected from the type of reference information, calibration by semantic structure-from-motion provides a remarkably larger number of reference points (more than factor 100). In tendency, the mean residuals of this method are slightly smaller (approx. one exponent), while

the standard deviations of the residuals are remarkably larger (approx. six exponents). These observations can be an indicator that the small amount of reference points in calibration by traffic signs fit well to each other, but may contain certain outliers in the observations that can't be uncovered by other observations. Second, the deviations of the interior orientation parameter values from reference calibration are remarkably smaller for calibration by semantic structure-from-motion (cf. scaling of vertical axes in Figures 8.11, 8.15, 8.16). Likewise, the precision of this method is remarkably better than the precision of calibration with traffic signs (cf. scaling of the vertical axis in Figures 8.11, 8.17, 8.18). Again, these observations are not surprising when comparing the number of reference points.

While the self-calibration methods are well suitable for doing camera calibration on the road, the experiments have confirmed that they are not suitable for analyzing the influence of the vehicle windshield. First, it is not possible to control the location and distribution of reference points in the image which is important to get estimates for the orientation parameters that are valid for the entire image area. Second, self-calibration has a larger number of error sources, which may interfere with a potential effect of the windshield on the estimates. For example, both erroneous vehicle positions from GPS and windshield refraction may contribute to a scale error that leads to an error in the estimated focal length. Furthermore, the experiments have revealed that pixel coordinates of the reference points extracted from traffic signs have worse accuracy than pixel coordinates extracted from well-tested and subpixel-accurate algorithms for circular reference marks or checkerboard patterns. Third and most important, it is practically not feasible to obtain pre-determined high-quality object coordinates for the reference points for self-calibration on the road which are needed to uncover a potential effect of the windshield, especially in the case of the large amount of reference points from image features. In contrast, the high effort for test field calibration shows that it can be used only under special conditions like in production plants or in research facilities, but not during regular drives of a car on public roads. While there are many differences between the proposed methods for self-calibration and test field calibration, the algorithms still have certain aspects in common. Especially the feature matching and 3d reconstruction steps of self-calibration by semantic structure-from-motion have several parallels to the matching and 3d reconstruction steps of the proposed test field calibration. Additionally, all proposed methods rely on bundle adjustment for final estimation of the orientation parameter values.

The results of the self-calibration methods and the test field calibration method are difficult to compare based on the estimated values, as different datasets have been used for the experiments and as the objectives of the methods are different. Furthermore, the method for test field calibration was not evaluated against a reference calibration. Therefore, only the residuals can be compared with each other, whereby the mean residuals of test field calibration are in tendency larger than the mean residuals of both self-calibration methods. Besides, the standard deviations of the residuals are similar to calibration by semantic structure-from-motion (based on the exponent, cf. Figures 8.1 and 8.2, Tables 8.4, 8.5, 8.7 and 8.8), but remarkably larger than the standard deviations for calibration with traffic signs. In general, it can be said that especially the large deviations of the estimated interior orientation parameter values for calibration with traffic signs from the reference calibration indicate a insufficient quality to analyze the influence of the windshield, as the windshield effect may cause deviations that are even smaller.

9 Conclusion

This chapter covers conclusions drawn from the work presented in this thesis (Section 9.1). Additionally, it covers future perspectives how the work could be continued (Section 9.2).

9.1 Conclusions on the research questions

The following conclusions correspond with the three research questions (Section 1.3).

9.1.1 Camera calibration with test fields

Q: How does a vehicle windshield in the optical path between a forward-looking on-board stereo camera system and a calibration test field influence the parameter values, standard deviations and correlations of the interior, relative and exterior orientation parameters estimated by test field calibration based on bundle adjustment in a set of experimental cases covering two kinds of test fields, different camera models and different parametrizations of stereo constraints?

A: By the proposed method, reference points from uncoded and coded reference marks on two kinds of non-rigid test fields have been jointly used to calibrate a stereo camera system. For the investigation of the windshield influence, calibration has been carried out successfully with and without a windshield in several experimental cases with different stereo constraints, camera models and reference points from either one or both test fields. This underlines the potential of test field calibration for investigations under specific conditions. Analysis has shown larger absolute residuals of the image points if a windshield is present, especially for experimental cases with a stereo constraint that estimates only the relative and the exterior orientation of one camera (cf. *Stereo constraint 1*). For most camera orientation parameters, statistically significant deviations have been obtained between calibration with and without the vehicle windshield. Larger deviations have been obtained for distortion parameters than for the focal length and principal point coordinates. Larger deviations have been obtained also for the camera model considering only radial distortions (cf. model *Radial*) instead of both radial and tangential distortions (*Both*). For most experimental cases, the precision of the estimated interior, relative and exterior orientation parameters has been better if no windshield was present. Correlation analysis has revealed small mean correlations indicating the absence of systematic dependencies between the orientation parameters in the imaging configuration. Remarkably, larger correlations have been achieved in average if no windshield was present. Calibration with a camera model without distortions (*None*) has failed. Calibration has also failed for few experimental cases if only the checkerboard test field together with a camera model considering only radial distortions was used. Finally, for many experimental cases, less different values have been obtained between cases where no stereo constraints are applied (*Stereo constraint 0*) and cases where the relative as well as the exterior orientation of both cameras are estimated during calibration (*Stereo constraint 2*) compared to cases with *Stereo constraint 1*. As conclusion from the previous statements, it can be recommended to use the camera model with both radial and tangential distortions, to use both test

fields (*Merged*) and stereo constraints where the radial and exterior orientation of both cameras are estimated.

9.1.2 Camera calibration with traffic signs

Q: Which types of traffic signs are most appropriate to derive reference points from by deep learning-based computer vision for self-calibration with a sequence of road scene images taken with a forward-looking on-board mono camera?

A: By the proposed method, calibration has been carried out successfully with traffic signs with triangle, rectangle or circle shape with different image sequences from road scenes, thereby showing the potential to calibrate automotive cameras by self-calibration on the road without a priori known reference information. Smaller deviations with regard to a reference calibration, performed as test field calibration, have been mostly obtained for experimental cases involving triangular traffic signs compared to rectangular or circular traffic signs. Smaller deviations have been also obtained when using *Deeplabv3+* for semantic segmentation compared to *EfficientPS*. In few experimental cases with rectangular traffic signs, the number of reference points that has been obtained is even lower than for single-image test field calibration and so calibration can't be considered as reliable in such cases. For the urban *Munich sequence*, a higher number of reference points and smaller deviations have been obtained compared to the suburban *Ettlingen sequence*. This can be interpreted as indicator that urban road scenes are more suitable for camera calibration with traffic signs than suburban scenes.

9.1.3 Camera calibration by semantic structure-from-motion

Q: How can semantic road scene knowledge and vehicle motion models be integrated into a structure-from-motion pipeline to improve self-calibration of a forward-looking on-board mono camera with a series of road scene images?

A: By the proposed method, camera calibration has been carried out successfully with image sequences from road scenes in different experimental cases using scene knowledge from semantic segmentation to make feature extraction and feature matching in the structure-from-motion pipeline more robust. Experimental evaluation has confirmed the theoretical expectation that using scene knowledge by means of exclusion masks during feature extraction leads to a smaller number of reference points compared to regular structure-from-motion, but also to smaller residuals (absolute mean values, standard deviations). Analysis of the deviations from reference test field calibration indicates a dependency of the estimated values and standard deviations of the interior orientation parameters on the set of semantic classes considered for creating the exclusion masks. Analysis supported by statistical significance tests has further revealed smaller deviations from reference calibration for the *Munich sequence* if scene knowledge has been used during feature matching compared to when it has been used not. It has also shown smaller deviations for the *Munich sequence* if a Kalman filter with a special vehicle motion model has been applied to refine the vehicle trajectory from GPS in order to achieve better metric scaling in the structure-from-motion pipeline. For calibration with the proposed method, generally a better precision of the interior orientation parameters has been obtained compared to the reference calibration. In comparison to self-calibration with traffic signs, both a remarkably higher number of reference points and remarkably lower deviations have been obtained with this method.

9.2 Future work

Based from the limitations of the proposed method, new possible directions for further investigations on the windshield influence could emerge from a modified experimental setup. By using a spare windshield in a lab environment instead of a real vehicle, the same test field orientations and so the same number and distribution of reference points can be obtained to achieve a better comparability between calibration with and without the windshield. Employing independent reference information that is not used for calibration, like separate objects with known shapes, would allow to determine the accuracy instead of only the precision of the estimated camera orientation parameters and thus obviously increase the value of the evaluation [Luhmann et al., 2016]. Further research for self-calibration with traffic signs could emerge from improving the deep networks for semantic segmentation, depth estimation and boundary detection. More sophisticated models that are currently published each year will very likely increase the performance of these methods with beneficial effects on camera calibration. Additionally, specialized networks could be employed in order to obtain a higher number of reference points. For example, classic image processing steps applied after the deep networks in the proposed method to determine the pixel and object coordinates of the reference points may be integrated into these networks. For the same reason, it seems obvious to add support for other traffic sign shapes, especially unique ones that can be easily identified in the images, like it is the case for the stop sign or the right of way sign. For calibration by semantic structure-from-motion, further research could address even more precise exclusion of undesired reference points by better scene knowledge: For example, it could be determined which individual vehicles are currently in motion or which parts of a building have problematic reflecting surfaces. Furthermore, a more sophisticated approach to select an initial image pair for 3d reconstruction could avoid non-converging bundle adjustments. Finally, especially for use in mass-produced vehicles, it can be interesting to determine an appropriate small number of images allowing calibration with sufficient quality. It can be also interesting to work on reducing the computational complexity of the proposed method, especially with regard to the global bundle adjustment. Thereby, it could be expected to better cope with limited hardware resources in a vehicle.

Bibliography

Abdel-Aziz YI, Karara H (1971) Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. Photogrammetric Engineering & Remote Sensing, 81 (2): 103–107.

Adamczyk M, Liberadzki P, Sitnik R (2018) Temperature compensation method for digital cameras in 2d and 3d measurement applications. Sensors, 18 (11): article no. 3685.

Agarwal S, Mierle K, Team TCS (2022) Ceres Solver. Software.

Albl C, Pajdla T (2014) Global camera parameterization for bundle adjustment. In: International Conference on Computer Vision Theory and Applications: 555–561.

Alvarez S, Llorca DF, Sotelo MA (2014) Hierarchical camera auto-calibration for traffic surveillance systems. Expert Systems with Applications, 41 (4/1): 1532–1542.

Armstrong M, Zisserman A, Hartley R (1996) Self-calibration from image triplets. In: Buxton B, Cipolla R (eds) European Conference on Computer Vision: Springer, LNCS, 1064, 1–16.

Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (12): 2481–2495.

Bao SY, Savarese S (2011) Semantic structure from motion. In: IEEE Conference on Computer Vision and Pattern Recognition: 2025–2032.

Barsan IA, Liu P, Pollefeys M, Geiger A (2018) Robust dense mapping for large-scale dynamic environments. In: International Conference on Robotics and Automation: 7510–7517.

Baumer GmbH (2019) VLG-20C.I. Technical specifications.

Bay H, Tuytelaars T, Van Gool L (2006) SURF: Speeded up robust features. In: Leonardis A, Bischof H, Pinz A (eds) European Conference on Computer Vision: Springer, LNCS, 3951, 404–417.

Bellino M, Holzmann F, Kolski S, de Meneses YL, Jacot J (2005) Calibration of an embedded camera for driver-assistant systems. In: IEEE Intelligent Transportation Systems Conference: 354–359.

Benenson R, Omran M, Hosang J, Schiele B (2015) Ten years of pedestrian detection, what have we learned? In: Agapito L, Bronstein MM, Rother C (eds) European Conference on Computer Vision Workshops: Springer, LNCS, 8926, 613–627.

Bergamasco F, Albarelli A, Rodolà E, Torsello A (2013) Can a fully unconstrained imaging model be applied effectively to central cameras? In: IEEE Conference on Computer Vision and Pattern Recognition: 1391–1398.

Bergmann P, Wang R, Cremers D (2018) Online photometric calibration of auto exposure video for realtime visual odometry and SLAM. IEEE Robotics and Automation Letters, 3 (2): 627–634.

Bertasius G, Shi J, Torresani L (2015) Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In: IEEE Conference on Computer Vision and Pattern Recognition: 4380–4389.

Bertozzi M, Bombini L, Broggi A, Grisleri P, Porta PP (2010) Camera-based automotive systems. In: Belbachir AN (ed) Smart Cameras: 319–338.

Bhardwaj R, Tummala GK, Ramalingam G, Ramjee R, Sinha P (2018) AutoCalib: Automatic traffic camera calibration at scale. ACM Transactions on Sensor Networks, 14 (3-4): article no. 19.

Bianco S, Ciocca G, Marelli D (2018) Evaluating the performance of structure from motion pipelines. Journal of Imaging, 4 (8): article no. 98.

Bjorkman M, Eklundh JO (2002) Real-time epipolar geometry estimation of binocular stereo heads. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (3): 425–432.

Bodis-Szomoru A, Daboczi T, Fazekas Z (2008) Calibration and sensitivity analysis of a stereo vision-based driver assistance system. In: Bhatti A (ed) Stereo Vision: 1–26.

Bogdan O, Eckstein V, Rameau F, Bazin JC (2018) DeepCalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In: ACM SIGGRAPH European Conference on Visual Media Production: article no. 6.

Borgmann B, Schatz V, Kieritz H, Scherer-Klöckling C, Hebel M, Arens M (2018) Data processing and recording using a versatile multi-sensor vehicle. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-1: 21–28.

Börlin N, Grussenmeyer P (2014) Camera calibration using the damped bundle adjustment toolbox. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, II-5: 89–96.

Broggi A, Bertozzi M, Fascioli A (2001) Self-calibration of a stereo vision system for automotive applications. In: International Conference on Robotics and Automation, 4, 3698–3703.

Brostow GJ, Shotton J, Fauqueur J, Cipolla R (2008) Segmentation and recognition using structure from motion point clouds. In: Forsyth D, Torr P, Zisserman A (eds) European Conference on Computer Vision: Springer, LNCS, 5302, 44–57.

Brown DC (1971) Close-range camera calibration. Photogrammetric Engineering, 37 (8): 855–866.

Brown LM, Fan Q, Zhai Y (2015) Self-calibration from vehicle information. In: IEEE International Conference on Advanced Video and Signal-Based Surveillance: 1–6.

Calonder M, Lepetit V, Ozuysal M, Trzcinski T, Strecha C, Fua P (2012) BRIEF: Computing a local binary descriptor very fast. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34 (7): 1281–1298.

Cannelle B, Paparoditis N, Pierrot-Deseilligny M, Papelard JP (2012) Off-line vs. on-line calibration of a panoramic-based mobile mapping system. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, I-3: 31–36.

Canny J (1986) A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8 (6): 679–698.

Casser V, Pirk S, Mahjourian R, Angelova A (2019) Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: AAAI Conference on Artificial Intelligence, 33 (1), 8001–8008.

Catala-Prat A, Rataj J, Reulke R (2006) Self-calibration system for the orientation of a vehicle camera. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVI-5: 68–73.

Chang H, Tsai F (2012) Reconstructing three-dimensional specific curve building models from a single perspective view image. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXIX-B6: 101–106.

Charles RQ, Su H, Kaichun M, Guibas LJ (2017) PointNet: Deep learning on point sets for 3d classification and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition: 77–85.

Chen L, Barron JT, Papandreou G, Murphy K, Yuille AL (2016) Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform. In: IEEE Conference on Computer Vision and Pattern Recognition: 4545–4554.

Chen L, Collins MD, Zhu Y, Papandreou G, Zoph B, Schroff F, Adam H, Shlens J (2018a) Searching for efficient multi-scale architectures for dense image prediction. In: Advances in Neural Information Processing Systems, 31.

Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018b) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) European Conference on Computer Vision: Springer, LNCS, 11211, 833–851.

Civera J, Gálvez-López D, Riazuelo L, Tardós JD, Montiel JMM (2011) Towards semantic SLAM using a monocular camera. In: IEEE/RSJ International Conference on Intelligent Robots and Systems: 1277–1284.

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The Cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition: 3213–3223.

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2019) Cityscapes dataset - benchmark suite - pixel-level semantic labeling task. Website. https://www.cityscapes-dataset.com/benchmarks/pixel-level-results, last accessed 2019-04-15.

Corres J, Jung J, Yoon I, Lee S, Paik J (2016) Object detection and tracking-based camera calibration for normalized human height estimation. Journal of Sensors, 2016: article no. 8347841.

Dang T, Hoffmann C, Stiller C (2009) Continuous stereo self-calibration by camera parameter tracking. IEEE Transactions on Image Processing, 18 (7): 1536–1550.

Deng R, Shen C, Liu S, Wang H, Liu X (2018) Learning to predict crisp boundaries. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) European Conference on Computer Vision: Springer, LNCS, 11210, 570–586.

Department of Transport - Ireland (2010) Traffic signs manual. Official regulations.

Dlugosz R, Dworakowski W, Suliga P (2019) Static camera calibration for advanced driver assistance system used in trucks - robust detector of calibration points. In: International Conference on Methods and Models in Automation and Robotics: 538–543.

Domhof J, Kooij JFP, Gavrila DM (2019) An extrinsic calibration tool for radar, camera and lidar. In: International Conference on Robotics and Automation: 8107–8113.

Dong J, Ge J, Luo Y (2007) Nighttime pedestrian detection with near infrared using cascaded classifiers. In: IEEE International Conference on Image Processing, 6, 185–188.

Dron L (1993) Dynamic camera self-calibration from controlled motion sequences. In: IEEE Conference on Computer Vision and Pattern Recognition: 501–506.

Dubey A (2016) Stereo vision - facing the challenges and seeing the opportunities for ADAS applications. Brochure.

Eigen D, Puhrsch C, Fergus R (2014) Depth map prediction from a single image using a multi-scale deep network. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ (eds) Advances in Neural Information Processing Systems: Curran Associates, 27.

El-Hakim S (1986) Real-time image metrology with CCD cameras. Photogrammetric Engineering & Remote Sensing, 52 (11): 1757–1766.

Elder J (2017) Determining rotations between disc axis and line of sight. Website. http://web.ncf.ca/aa456/scale/ellipse.html, last accessed 2017-10-29.

Engel J, Koltun V, Cremers D (2018) Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40 (3): 611–625.

Engel J, Schöps T, Cremers D (2014) LSD-SLAM: Large-scale direct monocular SLAM. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) European Conference on Computer Vision: Springer, LNCS, 8690, 834–849.

Enzweiler M, Gavrila DM (2009) Monocular pedestrian detection: Survey and experiments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31 (12): 2179–2195.

Faugeras OD, Luong QT, Maybank SJ (1992) Camera self-calibration: Theory and experiments. In: Sandini G (ed) European Conference on Computer Vision: Springer, LNCS, 588, 321–334.

Fischler MA, Bolles RC (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.

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

Franke U, Pfeiffer D, Rabe C, Knoeppel C, Enzweiler M, Stein F, Herrtwich RG (2013) Making bertha see. In: IEEE International Conference on Computer Vision Workshops: 214–221.

Fraser C (2013) Automatic camera calibration in close range photogrammetry. Photogrammetric Engineering & Remote Sensing, 79 (4): 381–388.

Fraser CS (1997) Digital camera self-calibration. ISPRS Journal of Photogrammetry and Remote Sensing, 52 (4): 149–159.

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB (2020) The MODISSA testbed. Website. https://www.iosb.fraunhofer.de/servlet/is/42840/, last accessed 2020-08-13.

Friel M, Savage DA, Hughes C, Ermilios P (2012) Online vehicle camera calibration based on road surface texture tracking and geometric properties. Patent. WO2012139636A1.

Fung GSK, Yung NHC, Pang GKH (2003) Camera calibration from road lane markings. Optical Engineering, 42 (10): 2967–2977.

Furukawa Y, Hernández C (2015) Multi-view stereo: A tutorial. Foundations and Trends in Computer Graphics and Vision, 9 (1-2): 1–148.

Garcia L (2017) Extrinsic calibration method for cameras of an on-board system for formation of stereo images. Patent. US10672147.

Garg R, B. G. VK, Carneiro G, Reid I (2016) Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: Leibe B, Matas J, Sebe N, Welling M (eds) European Conference on Computer Vision: Springer, LNCS, 9912, 740–756.

Ge J, Luo Y, Tei G (2009) Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. IEEE Transactions on Intelligent Transportation Systems, 10 (2): 283–298.

Gehrig S (2005) Large-field-of-view stereo for automotive applications. In: Workshop on Omnidirectional Vision, Camera Networks and Nonclassical Cameras

Geiger A, Moosmann F, Car Ö, Schuster B (2012) Automatic camera and range sensor calibration using a single shot. In: International Conference on Robotics and Automation: 3936–3943.

Ghasemi A, Zahediasl S (2012) Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, 10 (2): 486–489.

Gil G, Savino G, Piantini S, Pierini M (2018a) Motorcycle that see: Multifocal stereo vision sensor for advanced safety systems in tilting vehicles. Sensors, 18 (1): article no. 295.

Gil G, Savino G, Piantini S, Pierini M (2018b) Satellite markers: a simple method for ground truth car pose on stereo video. In: Verikas A, Radeva P, Nikolaev D, Zhou J (eds) International Conference on Machine Vision: SPIE, 10696, article no. 1069620.

Gil Y, Elmalem S, Haim H, Marom E, Giryes R (2019) MonSter: Awakening the mono in stereo. Computing Research Repository, arXiv preprint arXiv:1910.13708.

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition: 580–587.

Godard C, Aodha OM, Firman M, Brostow G (2019) Digging into self-supervised monocular depth estimation. In: IEEE/CVF International Conference on Computer Vision: 3827–3837.

Gopaul NS, Wang J, Hu B (2016) Camera auto-calibration in GPS/INS/stereo camera integrated kinematic positioning and navigation system. The Journal of Global Positioning Systems, 14 (1): article no. 3.

Graeter J, Wilczynski A, Lauer M (2018) LIMO: Lidar-monocular visual odometry. In: IEEE/RSJ International Conference on Intelligent Robots and Systems: 7872–7879.

Granshaw SI (2020) Photogrammetric terminology: fourth edition. The Photogrammetric Record, 35 (170): 143–288.

Guerrero JJ, Martinez-Cantin R, Sagüés C (2005) Visual map-less navigation based on homographies. Journal of Field Robotics, 22: 569–581.

Guindel C, Beltrán J, Martín D, García F (2017) Automatic extrinsic calibration for lidar-stereo vehicle sensor setups. In: IEEE International Conference on Intelligent Transportation Systems: 1–6.

Guo-Qing Wei, Song De Ma (1994) Implicit and explicit camera calibration: theory and experiments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16 (5): 469–480.

Häne C, Heng L, Lee GH, Fraundorfer F, Furgale P, Sattler T, Pollefeys M (2017) 3d visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. Image and Vision Computing, 68: 14–27.

Hanel A, Hoegner L, Stilla U (2016) Towards the influence of a car windshield on depth calculation with a stereo camera system. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLI-B5: 461–468.

Hanel A, Kreuzpaintner D, Stilla U (2018) Evaluation of a traffic sign detector by synthetic image data for advanced driver assistance systems. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2: 425–432.

Hanel A, Stilla U (2017) Structure-from-motion for calibration of a vehicle camera system with non-overlapping fields-of-view in an urban environment. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-1/W1: 181–188.

Hanel A, Stilla U (2018) Iterative calibration of a vehicle camera using traffic signs detected by a convolutional neural network. In: International Conference on Vehicle Technology and Intelligent Transport Systems: 187–195.

Hanel A, Stilla U (2019a) Evaluation of iterative calibration of vehicle cameras using reference information from traffic signs. In: Donnellan B, Klein C, Helfert M, Gusikhin O (eds) Smart Cities, Green Technologies and Intelligent Transport Systems.: Springer, CCIS, 992, 244–265.

Hanel A, Stilla U (2019b) Semantic road scene knowledge for robust self-calibration of environment-observing vehicle cameras. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W16: 103–110.

Hanel A, Sudi P, Pfenninger S, Steinbach E, Stilla U (2019) Filter-based pose estimation for electric vehicles relative to a ground-based charging platform using on-board camera images. In: Kersten TP (ed) Wissenschaftlich-Technische Jahrestagung der DGPF, 28, 54–67.

Hansen P, Alismail H, Rander P, Browning B (2012) Online continuous stereo extrinsic parameter estimation. In: IEEE Conference on Computer Vision and Pattern Recognition: 1059–1066.

Hariharan B, Arbeláez P, Girshick R, Malik J (2014) Simultaneous detection and segmentation. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) European Conference on Computer Vision: Springer, LNCS, 8695, 297–312.

Hartley R, Zisserman A (2003) Multiple View Geometry in Computer Vision. Cambridge University Press.

Hartley RI (1993) Euclidean reconstruction from uncalibrated views. In: Applications of Invariance in Computer Vision: 235–256.

Hartley RI (1997a) Lines and points in three views and the trifocal tensor. International Journal of Computer Vision, 22 (2): 125–140.

Hartley RI (1997b) Self-calibration of stationary cameras. International Journal of Computer Vision, 22 (1): 5–23.

Hastedt H, Ekkel T, Luhmann T (2016) Evaluation of the quality of action cameras with wide-angle lenses in UAV photogrammetry. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLI-B1: 851–859.

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: IEEE International Conference on Computer Vision: 2980–2988.

Heikkila J (2000) Geometric camera calibration using circular control points. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (10): 1066–1077.

Heikkila J, Silven O (1997) A four-step camera calibration procedure with implicit image correction. In: IEEE Conference on Computer Vision and Pattern Recognition: 1106–1112.

Hella Gutmann Solutions GmbH (2016) CSC-Tool. Operating Instructions. Manual.

Hella Gutmann Solutions GmbH (2019) Image of the CSC calibration test field. Website. Website, https://www.hella-gutmann.com/fileadmin/00_HGS_Bilder/A_Workshopsolutions/A3_Pruef-_u_Einstellwerkzeuge/A3_1_CSC-Tool/003_hgs_csc_ansicht_gr.png, last accessed 2019-09-13.

Heng L, Bürki M, Lee GH, Furgale P, Siegwart R, Pollefeys M (2014) Infrastructure-based calibration of a multi-camera rig. In: International Conference on Robotics and Automation: 4912–4919.

Heng L, Lee GH, Pollefeys M (2015) Self-calibration and visual SLAM with a multi-camera system on a micro aerial vehicle. Autonomous Robots, 39 (3): 259–277.

Heng L, Li B, Pollefeys M (2013) CamOdoCal: Automatic intrinsic and extrinsic calibration of a rig with multiple generic cameras and odometry. In: IEEE/RSJ International Conference on Intelligent Robots and Systems: 1793–1800.

Henry P, Krainin M, Herbst E, Ren X, Fox D (2012) RGB-D mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. The International Journal of Robotics Research, 31 (5): 647–663.

Herrmann C, Ruf M, Beyerer J (2018) CNN-based thermal infrared person detection by domain adaptation. In: Dudzik MC, Ricklin JC (eds) Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything: SPIE, 10643, article no. 1064308.

Heyden A, Astrom K (1996) Euclidean reconstruction from constant intrinsic parameters. In: International Conference on Pattern Recognition, 1, 339–343.

Heyden A, Astrom K (1997) Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. In: IEEE Conference on Computer Vision and Pattern Recognition: 438–443.

Hirzer M, Roth PM, Lepetit V (2017) Efficient 3d tracking in urban environments with semantic segmentation. In: Kim TK, Zafeiriou S, Brostow G, Mikolajczyk K (eds) British Machine Vision Conference: 143.1–143.12.

Hold S, Gormer S, Kummert A, Meuter M, Muller-Schneiders S (2009) A novel approach for the online initial calibration of extrinsic parameters for a car-mounted camera. In: IEEE International Conference on Intelligent Transportation Systems: 1–6.

Hold-Geoffroy Y, Sunkavalli K, Eisenmann J, Fisher M, Gambaretto E, Hadap S, Lalonde J (2018) A perceptual measure for deep single image camera calibration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2354–2363.

Houben S (2014) Towards the intrinsic self-calibration of a vehicle-mounted omni-directional radially symmetric camera. In: IEEE Intelligent Vehicles Symposium: 878–883.

Houben S, Stallkamp J, Salmen J, Schlipsing M, Igel C (2013) Detection of traffic signs in real-world images: The German traffic sign detection benchmark. In: International Joint Conference on Neural Networks: 1–8.

Huang R, Ye Z, Hong D, Xu Y, Stilla U (2019) Semantic labeling and refinement of lidar point clouds using deep neural network in urban areas. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-2/W7: 63–70.

Ishikawa R, Oishi T, Ikeuchi K (2018) Offline and online calibration of mobile robot and SLAM device for navigation. Computing Research Repository, arXiv preprint arXiv:1804.04817.

Ismail K, Sayed T, Saunier N (2010) Camera calibration for urban traffic scenes: Practical issues and a robust approach. In: Transportation Research Board Annual Meeting Compendium of Papers

Janai J, Güney F, Behl A, Geiger A (2017) Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. Computing Research Repository, arXiv preprint arXiv:1704.05519.

Jiang C, Paudel DP, Fougerolle Y, Fofi D, Demonceaux C (2016) Static-map and dynamic object reconstruction in outdoor scenes using 3-d motion segmentation. IEEE Robotics and Automation Letters, 1 (1): 324–331.

Kahmen O, Rofallski R, Luhmann T (2020) Impact of stereo camera calibration to object accuracy in multimedia photogrammetry. Remote Sensing, 12 (12): article no. 2057.

Kalman RE (1960) A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82 (1): 35–45.

Kaneko M, Iwami K, Ogawa T, Yamasaki T, Aizawa K (2018) Mask-SLAM: Robust feature-based monocular SLAM by masking using semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops: 371–379.

Kannala J, Brandt SS (2006) A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (8): 1335–1340.

Keivan N, Sibley G (2015) Online SLAM with any-time self-calibration and automatic change detection. In: International Conference on Robotics and Automation: 5775–5782.

Keller CG, Enzweiler M, Rohrbach M, Llorca DF, Schnorr C, Gavrila DM (2011) The benefits of dense stereo for pedestrian detection. IEEE Transactions on Intelligent Transportation Systems, 12 (4): 1096–1106.

Kluger F, Ackermann H, Yang MY, Rosenhahn B (2017) Deep learning for vanishing point detection using an inverse gnomonic projection. In: Roth V, Vetter T (eds) German Conference on Pattern Recognition: Springer, LNCS, 10496, 17–28.

Knorr M (2018) Self-Calibration of Multi-Camera Systems for Vehicle Surround Sensing. PhD thesis, Karlsruhe Institude of Technology, Department of Mechanical Engineering, Institute of Measurement and Control Systems.

Kraus K (2007) Photogrammetry - Geometry from Images and Laser Scans. De Gruyter.

Kruger LE, Wohler C, Wurz-Wessel A, Stein F (2004) In-factory calibration of multiocular camera systems. In: Osten W, Takeda M (eds) Optical Metrology in Production Engineering: SPIE, 5457, 126–137.

Labatut P, Pons JP, Keriven R (2007) Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In: IEEE International Conference on Computer Vision: 1–8.

Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N (2016) Deeper depth prediction with fully convolutional residual networks. In: International Conference on 3D Vision: 239–248.

Lamprecht B, Rass S, Fuchs S, Kyamakya K (2007) Extrinsic camera calibration for an on-board two-camera system without overlapping field of view. In: IEEE International Conference on Intelligent Transportation Systems: 265–270.

Lasaruk A, Hachfeld F (2019) Method and apparatus for the autocalibration of a vehicle camera system. Patent. US20190297314A1.

Lasaruk A, Neralla DF (2018) Method and apparatus for the compensation of static image distortions introduced by a windshield onto an adas camera. Patent. EP3293701A1.

Lee GH, Faundorfer F, Pollefeys M (2013) Motion estimation for self-driving cars with a generalized camera. In: IEEE Conference on Computer Vision and Pattern Recognition: 2746–2753.

Leite N, Bue AD, Gaspar J (2008) Calibrating a network of cameras - based on visual odometry. In: Jornadas de Engenharia Electronica e Telecomunicacoes e de Computadores: 174–179.

Levinson J, Thrun S (2013) Automatic online calibration of cameras and lasers. In: Robotics: Science and Systems

Leys C, Ley C, Klein O, Bernard P, Licata L (2013) Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49 (4): 764–766.

Li C, Lin S, Zhou K, Ikeuchi K (2017) Radiometric calibration from faces in images. In: IEEE Conference on Computer Vision and Pattern Recognition: 1695–1704.

Li X, Belaroussi R (2016) Semi-dense 3d semantic mapping from monocular SLAM. Computing Research Repository, arXiv preprint arXiv:1611.04144.

Liebowitz D (2001) Camera Calibration and Reconstruction of Geometry from Images. PhD thesis, University of Oxford, Department of Engineering Science, Robotics Research Group.

Lin TY, Dollár P, Girshick RB, He K, Hariharan B, Belongie SJ (2017) Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition: 936–944.

Ling Y, Shen S (2016) High-precision online markerless stereo extrinsic calibration. In: IEEE/RSJ International Conference on Intelligent Robots and Systems: 1771–1778.

Liu S, Ding W, Liu C, Liu Y, Wang Y, Li H (2018) ERN: Edge loss reinforced semantic segmentation network for remote sensing images. Remote Sensing, 10 (9): article no. 1339.

Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu C, Berg AC (2016) SSD: Single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M (eds) European Conference on Computer Vision: Springer, LNCS, 9905, 21–37.

Liu Y, Cheng MM, Hu X, Bian JW, Zhang L, Bai X, Tang J (2019) Richer convolutional features for edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41 (8): 1939–1946.

Livyatan H, Berberian O (2017) Stereo auto-calibration from structure-from-motion. Patent. US20170019657A1.

Lowe DG (1999) Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision, 2, 1150–1157.

Lu X, Wang Y, Ling Z, Wang K, Wang G (2013) A method for vehicle-mounted camera calibration under urban traffic scenes. In: Chinese Automation Congress: 556–560.

Luhmann T, Fraser C, Maas HG (2016) Sensor modelling and camera calibration for close-range photogrammetry. ISPRS Journal of Photogrammetry and Remote Sensing, 115: 37–46.

Luhmann T, Robson S, Kyle S, Boehm J (2013) Close-range Photogrammetry and 3D Imaging. De Gruyter.

Luhmann T, Robson S, Kyle S, Harley I (2006) Close Range Photogrammetry. Principles, Methods and Applications. Whittles Publishing.

Maas HG (2015a) A modular geometric model for underwater photogrammetry. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-5/W5: 139–141.

Maas HG (2015b) On the accuracy potential in underwater/multimedia photogrammetry. Sensors, 15 (8): 18140–18152.

Mahe H, Marraud D, Comport AI (2018) Semantic-only visual odometry based on dense class-level segmentation. In: International Conference on Pattern Recognition: 1989–1995.

Marita T, Oniga F, Nedevschi S, Graf T, Schmidt R (2006) Camera calibration method for far range stereovision sensors used in vehicles. In: IEEE Intelligent Vehicles Symposium: 356–363.

Marr D, Hildreth E (1980) Theory of edge detection. Proceedings of the Royal Society of London: Biological sciences, 207 (1167): 187–217.

Massey FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association, 46 (253): 68–78.

Maybank SJ, Faugeras OD (1992) A theory of self-calibration of a moving camera. International Journal of Computer Vision, 8 (2): 123–151.

Mentzer N, Mahr J, Payá-Vayá G, Blume H (2019) Online stereo camera calibration for automotive vision based on HW-accelerated A-KAZE-feature extraction. Journal of Systems Architecture, 97: 335–348.

Mentzer N, Vayá G, Blume H, von Egloffstein N, Krueger L (2017) Self-calibration of wide baseline stereo camera systems for automotive applications. In: Guillermo Payá-Vayá HB (ed) Towards a Common Software/Hardware Methodology for Future Advanced Driver Assistance Systems: 157–200.

Mertan A, Duff DJ, Unal G (2022) Single image depth estimation: An overview. Digital Signal Processing, 123: article no. 103441.

Miksch M, Yang B, Zimmermann K (2010) Homography-based extrinsic self-calibration for cameras in automotive applications. In: Workshop on Intelligent Transportation

Mitishita E, Cortes J, Centeno J, Machado A (2009) Small-format digital camera: A study into stability analysis of the interior orientation parameters through temperature variation. In: International Symposium on Mobile Mapping Technology: 121–134.

Mo Z, Shi B, Yeung S, Matsushita Y (2017) Radiometric calibration for internet photo collections. In: IEEE Conference on Computer Vision and Pattern Recognition: 275–283.

Mohan R, Valada A (2021) EfficientPS: Efficient panoptic segmentation. International Journal of Computer Vision, 129 (5): 1551–1579.

Mueller GR, Wuensche H (2017) Continuous stereo camera calibration in urban scenarios. In: IEEE International Conference on Intelligent Transportation Systems: 1–6.

Muhovic J, Pers J (2020) Correcting decalibration of stereo cameras in self-driving vehicles. Sensors, 20 (11): article no. 3241.

Mur-Artal R, Montiel JMM, Tardós JD (2015) ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics, 31 (5): 1147–1163.

Mur-Artal R, Tardós JD (2017) ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. IEEE Transactions on Robotics, 33 (5): 1255–1262.

Murali V, Chiu H, Samarasekera S, Kumar RT (2017) Utilizing semantic visual landmarks for precise vehicle navigation. In: IEEE International Conference on Intelligent Transportation Systems: 1–8.

Musleh B, Martín D, Armingol JM, de la Escalera A (2014) Extrinsic parameter self-calibration and non-linear filtering for in-vehicle stereo vision systems at urban environments. In: International Conference on Computer Vision Theory and Applications: 427–434.

Nießner M, Zollhöfer M, Izadi S, Stamminger M (2013) Real-time 3d reconstruction at scale using voxel hashing. ACM Transactions on Graphics, 32 (6): 1–11.

Nister D, Naroditsky O, Bergen J (2006) Visual odometry for ground vehicle applications. Journal of Field Robotics, 23: 3–20.

Okouneva G (2017) Vehicle vision system with targetless camera calibration. Patent. US9563951B2.

Onsemi (2017) AR0331 1/3-Inch 3.1 Mp/FullHD Digital Image Sensor. Semiconductor Components Industries. Datasheet. Publication order number AR0331/D.

OpenCV (2017) Camera calibration and 3d reconstruction. Website. http://opencv.org, last accessed 2017-01-30.

Pagel F, Willersinn D (2011) Extrinsic camera calibration in vehicles with explicit ground estimation. In: Workshop on Intelligent Transportation

Paone JR, Karnowski T, Aykac D, Ferrell R, Goddard Jr J, Albright AP (2019) Investigating camera calibration methods for naturalistic driving studies. In: IS&T International Symposium on Electronic Imaging

Paula MBD, Jung CR, Silveira LGD (2014) Automatic on-the-fly extrinsic camera calibration of onboard vehicular cameras. Expert Systems with Applications, 41 (4/2): 1997–2007.

Pekkucuksen IE, Batur AU (2018) Method, apparatus and system for performing geometric calibration for surround view camera solution. Patent. US9892493B2.

Pflug C, Platonov J, Kaczmarczyk P, Gebauer T (2013) Method and device for online calibration of vehicle cameras. Patent. WO2013107814A1.

Pliefke S (2013) Calibration system and method for vehicular surround vision system. Patent. WO2013074604A2.

Polic M, Förstner W, Pajdla T (2018) Fast and accurate camera covariance computation for large 3d reconstruction. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) European Conference on Computer Vision: Springer, LNCS, 11206, 697–712.

Pollefeys M, Van Gool L (1997) Self-calibration from the absolute conic on the plane at infinity. In: Computer Analysis of Images and Patterns: 175–182.

Qi X, Yang S, Yan Y (2018) Deep learning based semantic labelling of 3d point cloud in visual SLAM. IOP Conference Series: Materials Science and Engineering, 428: article no. 012023.

Redmon J, Farhadi A (2017) YOLO9000: Better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition: 6517–6525.

Rehder E, Kinzig C, Bender P, Lauer M (2017) Online stereo camera calibration from scratch. In: IEEE Intelligent Vehicles Symposium: 1694–1699.

Remondino F, Fraser C (2006) Digital camera calibration methods: considerations and comparisons. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVI-5: 266–272.

Remondino F, Nocerino E, Toschi I, Menna F (2017) A critical review of automated photogrammetric processing of large datasets. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W5: 591–599.

Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (6): 1137–1149.

Ribeiro AAGA, Dihl LL, Jung CR (2006) Automatic camera calibration for driver assistance systems. In: International Conference on Systems, Signals and Image Processing: 173–176.

Robert Bosch GmbH (2018) Advanced driver assistance system - precise and efficient adjustment with Bosch. Brochure.

Rodehorst V, Heinrichs M, Hellwich O (2008) Evaluation of relative pose estimation methods for multi-camera setups. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVII-B3b: 135–140.

Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical Image Computing and Computer-Assisted Intervention: Springer, LNCS, 9351, 234–241.

Rosebrock D, Wahl FM (2012) Generic camera calibration and modeling using spline surfaces. In: IEEE Intelligent Vehicles Symposium: 51–56.

Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: An efficient alternative to SIFT or SURF. In: IEEE International Conference on Computer Vision: 2564–2571.

Ruland T, Loose H, Pajdla T, Krüger L (2010) Hand-eye autocalibration of camera positions on vehicles. In: IEEE International Conference on Intelligent Transportation Systems: 367–372.

Rünz M, Agapito L (2017) Co-Fusion: Real-time segmentation, tracking and fusion of multiple objects. In: International Conference on Robotics and Automation: 4471–4478.

Runz M, Buffier M, Agapito L (2018) MaskFusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In: IEEE International Symposium on Mixed and Augmented Reality: 10–20.

Scaramuzza D, Fraundorfer F (2011) Visual odometry [tutorial]. IEEE Robotics & Automation Magazine, 18 (4): 80–92.

Scaramuzza D, Fraundorfer F, Pollefeys M, Siegwart R (2009) Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In: IEEE International Conference on Computer Vision: 1413–1419.

Scaramuzza D, Martinelli A, Siegwart R (2006) A flexible technique for accurate omnidirectional camera calibration and structure from motion. In: IEEE International Conference on Computer Vision Systems

Scheller S, Westfeld P, Ebersbach D (2007) Calibration of a mobile mapping camera system with photogrammetric methods. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVI-5/C55.

Schneider CT, Bösemann W, Godding R (2017) AICON 3D Systems GmbH. Software Aicon 3D Studio.

Schöller C, Schnettler M, Krämmer A, Hinz G, Bakovic M, Güzet M, Knoll A (2019) Targetless rotational auto-calibration of radar and camera for intelligent transportation systems. In: IEEE International Conference on Intelligent Transportation Systems: 3934–3941.

Schönberger J, Pollefeys M, Geiger A, Sattler T (2018) Semantic visual localization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition: 6896–6906.

Schönberger JL, Frahm J (2016) Structure-from-motion revisited. In: IEEE Conference on Computer Vision and Pattern Recognition: 4104–4113.

Schubert R, Richter E, Wanielik G (2008) Comparison and evaluation of advanced motion models for vehicle tracking. In: International Conference on Information Fusion: 1–6.

Sermanet P, LeCun Y (2011) Traffic sign recognition with multi-scale convolutional networks. In: International Joint Conference on Neural Networks: 2809–2813.

Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (4): 640–651.

Shih SW, Hung YP, Lin WS (1996) Accuracy analysis on the estimation of camera parameters for active vision systems. In: International Conference on Pattern Recognition, 1, 930–935.

Stamatopoulos C, Fraser CS (2014) Automated target-free network orientation and camera calibration. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, II-5: 339–346.

Stein GP, Gdalyahu Y, Shashua A (2010) Stereo-Assist: Top-down stereo for driver assistance systems. In: IEEE Intelligent Vehicles Symposium: 723–730.

Stueckler J, Biresev N, Behnke S (2012) Semantic mapping using object-class segmentation of RGB-D images. In: IEEE/RSJ International Conference on Intelligent Robots and Systems: 3005–3010.

Sturm PF, Maybank SJ (1999) On plane-based camera calibration: A general algorithm, singularities, applications. In: IEEE Conference on Computer Vision and Pattern Recognition, 1, 432–437.

Sünderhauf N, Pham TT, Latif Y, Milford M, Reid I (2017) Meaningful maps with object-oriented semantic mapping. In: IEEE/RSJ International Conference on Intelligent Robots and Systems: 5079–5085.

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

SVS-Vistek GmbH (2020) Camera eco655MVGE. Datasheet. https://www.svs-vistek.com/de/svcam-kameras/svs-svcam-factsheet.php?id=eco655MVGE&type=cameras&lang=de, last accessed 2020-08-13.

Szczepanski M (2019) Online stereo camera calibration on embedded systems. PhD thesis, Universite Clermont Auvergne.

Tan J, Li J, An X, He H (2011) An interactive method for extrinsic parameter calibration of onboard camera. In: IEEE Intelligent Vehicles Symposium: 236–241.

Tareen SAK, Saleem Z (2018) A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. In: International Conference on Computing, Mathematics and Engineering Technologies: 1–10.

Texa (2017) ADAS solutions - maintenance of advanced driver assistance systems. Technical report.

Texa (2022) All around calibration system. Image. https://www.texa.com/wp-content/uploads/2021/04/Acc_ACS.jpg, last accessed 2022-07-04.

Thatcham Research and ADAS Repair Group (2016) Code of Practice For the Replacement & Refitting of Automotive Glazing for vehicles fitted with screen mounted Advanced Driver Assistance Systems (ADAS). Manual.

Triggs B (1997) Autocalibration and the absolute quadric. In: IEEE Conference on Computer Vision and Pattern Recognition: 609–614.

Triggs B, McLauchlan PF, Hartley RI, Fitzgibbon AW (2000) Bundle adjustment - a modern synthesis. In: Vision Algorithms: Theory and Practice: 298–372.

Tsai R (1987) A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf TV cameras and lenses. IEEE Journal on Robotics and Automation, 3 (4): 323–344.

Urban S, Leitloff J, Hinz S (2015) Improved wide-angle, fisheye and omnidirectional camera calibration. ISPRS Journal of Photogrammetry and Remote Sensing, 108: 72–79.

Vedaldi A, Guidi G, Soatto S (2007) Moving forward in structure from motion. In: IEEE Conference on Computer Vision and Pattern Recognition: 1–7.

Verbiest F, Proesmans M, Van Gool L (2020) Modeling the effects of windshield refraction for camera calibration. In: Vedaldi A, Bischof H, Brox T, Frahm JM (eds) European Conference on Computer Vision: Springer, LNCS, 12351, 397–412.

Vertens J, Valada A, Burgard W (2017) SMSnet: Semantic motion segmentation using deep convolutional neural networks. In: IEEE/RSJ International Conference on Intelligent Robots and Systems: 582–589.

Vo MN, Wang Z, Luu L, Ma J (2011) Advanced geometric camera calibration for machine vision. Optical Engineering, 50 (11): 1–4.

VS Technology (2015) Factsheet Optics SV-0614H. Website. https://www.vst.co.jp/de/products/machinevision/lenses/mega-pixel-cctv-lenses/, last accessed 2015-12-08.

Wang K, Lin Y, Wang L, Han L, Hua M, Wang X, Lian S, Huang B (2018) A unified framework for mutual improvement of SLAM and semantic segmentation. Computing Research Repository, arXiv preprint arXiv:1812.10016.

Wang R, Schwörer M, Cremers D (2017) Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras. In: IEEE International Conference on Computer Vision: 3923–3931.

Welch BL (1947) The generalisation of student's problems when several different population variances are involved. Biometrika, 34 (1-2): 28–35.

Winner H, Hakuli S, Lotz F, Singer C, eds (2015) Handbook of Driver Assistance Systems: Basic Information, Components and Systems for Active Safety and Comfort. Springer.

Wu C (2013) Towards linear-time incremental structure from motion. In: International Conference on 3D Vision: 127–134.

Younes G, Asmar D, Zelek J (2019) FDMO: Feature assisted direct monocular odometry. In: International Conference on Computer Vision Theory and Applications: 737–747.

Yousif K, Bab-Hadiashar A, Hoseinnezhad R (2015) An overview to visual odometry and visual SLAM: Applications to mobile robotics. Intelligent Industrial Systems, 1 (4): 289–311.

Yu C, Liu Z, Liu XJ, Xie F, Yang Y, Wei Q, Fei Q (2018) DS-SLAM: A semantic visual SLAM towards dynamic environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems: 1168–1174.

Yu Z, Feng C, Liu MY, Ramalingam S (2017) CASENet: Deep category-aware semantic edge detection. In: IEEE Conference on Computer Vision and Pattern Recognition: 1761–1770.

Zabatani A, Bareket S, Menashe O, Sperling E, Bronstein A, Bronstein M, Kimmel R, Surazhsky V (2017) Online compensation of thermal distortions in a stereo depth camera. Patent. US20170094255A1.

Zhang B, Appia V, Pekkucuksen I, Liu Y, Batur AU, Shastry P, Liu S, Sivasankaran S, Chitnis K (2014) A surround view camera solution for embedded systems. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops: 676–681.

Zhang Q, Pless R (2004) Extrinsic calibration of a camera and laser range finder (improves camera calibration). In: IEEE/RSJ International Conference on Intelligent Robots and Systems, 3, 2301–2306.

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

Zheng Y, Zhao W (2017) Can vehicle become a new pattern for roadside camera calibration? In: Chen CS, Lu J, Ma KK (eds) Asian Conference On Computer Vision Workshops: Springer, LNCS, 10117, 175–188.

Zhou Z, Farhat F, Wang JZ (2017) Detecting dominant vanishing points in natural scenes with application to composition-sensitive image retrieval. IEEE Transactions on Multimedia, 19 (12): 2651–2665.

Zhu Z, Liang D, Zhang S, Huang X, Li B, Hu S (2016) Traffic-sign detection and classification in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition: 2110–2118.

Ziebinski A, Cupek R, Erdogan H, Waechter S (2016) A survey of ADAS technologies for the future perspective of sensor fusion. In: Nguyen NT, Iliadis L, Manolopoulos Y, Trawiński B (eds) Computational Collective Intelligence: Springer, LNCS, 9876, 135–146.

Zou W, Li S (2015) Calibration of nonoverlapping in-vehicle cameras with laser pointers. IEEE Transactions on Intelligent Transportation Systems, 16 (3): 1348–1359.

Acknowledgment

Working on this thesis was a great opportunity for me to enhance both my professional and personal skills. Therefore, I am very glad that I had this option. First of all, I would like to thank Prof. Uwe Stilla for offering me the freedom to do research on a topic of my own choice, for his support during my work and for the numerous opportunities to present it at national and international conferences. In particular, I am grateful for his valuable feedback and the continuous encouragement to push the dissertation forward. Second, I would like to thank Prof. Boris Jutzi from the Karlsruhe Institute of Technology for being the second reviewer of my thesis. His suggestions and recommendations towards finalization of the dissertation were very fruitful and allowed me to further improve it. In addition, I would like to thank Prof. Urs Hugentobler for being the chairman of the examination committee.

I would also like to thank several other people who contributed to drive my work forward. First, I would like to mention the colleagues from the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, Ettlingen, in particular Dr. Marcus Hebel, Dr. Björn Borgmann and Dr. Joachim Gehrung for acquiring and providing the test data for several research questions addressed in this thesis. Without their support, tests of the proposed methods would have been heavily limited. I would also like to mention my former chair colleague Dr. Sebastian Tuttas for helping me to acquire the other test data used for this thesis. Furthermore, I would like to thank my former chair colleague Prof. Ludwig Hoegner for his support in many questions concerning both my research and teaching activities. Moreover, I would like to express my gratitude to the colleagues at the Unit for Photogrammetry and Remote Sensing and the Chair of Remote Sensing Technology at the Technical University of Munich for both fruitful professional discussions and a pleasant personal time spent together. Also, I would like to thank several students who have contributed to this thesis.

Last, but most certainly not least, I would like to thank my family, relatives and friends for their support. Above all, I would like to mention my parents Gertraud and Günter for always covering my back what allowed me to work towards my goals.