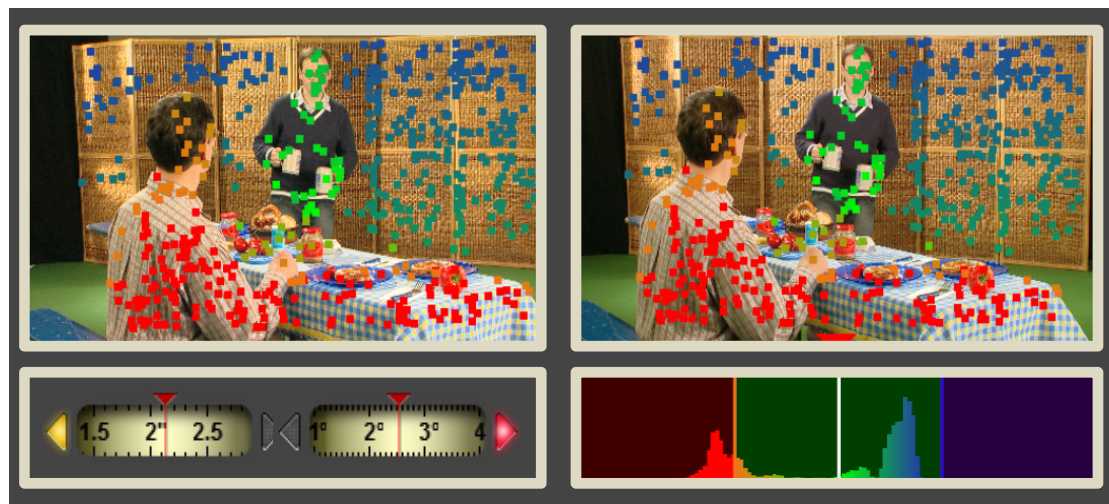


Method for the Automated Analysis, Control and Correction of Stereoscopic Distortions and Parameters for 3D-TV Applications

New Image Processing Algorithms to Improve the Efficiency of Stereo- and Multi-Camera 3D-TV Productions



Frederik Zilly

Method for the Automated Analysis, Control and Correction of Stereoscopic Distortions and Parameters for 3D-TV Applications

New Image Processing Algorithms to Improve the Efficiency of Stereo- and Multi-Camera 3D-TV Productions

vorgelegt von
Dipl.-Phys.
Frederik Zilly
geb. in Fortaleza/Brasilien

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr.-Ing. Sebastian Möller
Gutachter:	Prof. Dr.-Ing. Thomas Sikora
	Prof. Dr.-Ing. Peter Eisert (HU Berlin)
	Prof. Dr.-Ing. Olaf Hellwich

Tag der wissenschaftlichen Aussprache: 28. Juli 2015

Berlin 2015

Abstract

The background and motivation for the research performed within this thesis is the introduction of the Digital Cinema which allows for new workflows based on image processing algorithms. Thereby, the development of algorithms for stereoscopic 3D and multi-camera productions within the era of the Digital Cinema is of special interest.

Several 3D productions have been released in the cinemas in the past years¹ while the basic principle of 3D reproduction is still based on Wheatstone's [Wheatstone38] and Brewster's stereoscopic approach [Brewster56] where two views corresponding to two different viewing positions are presented to the viewer's left and right eye. However, if the reproduced 3D content imposes unnatural viewing conditions when watched, e.g. due to an excessive amount of inherent parallax, an impaired 3D sensation can result which can even lead to visual fatigue and head-ache [IJsselsteijn00]. Consequently, specific 3D production rules as described in [Mendiburu08] and [Knorr12] have to be obeyed when high quality 3D content shall be produced. It includes a precise calibration of the two cameras with consistent electronic and optical parameters. Moreover, the stereo baseline and convergence distance have to be chosen according to the depth structure of the scene content. When performed without specific assistance systems, the calibration process and the choice of proper stereoscopic parameters as described by Lipton in [Lipton82] can be tedious tasks which require trained personnel and increase the overall production costs [Buchs11]. With the advent of digital cameras, it became possible to analyze and possibly correct the 3D signal electronically using dedicated stereoscopic image processors [Zilly10b, Sony] which facilitates the above mentioned tasks and allows for new 3D production workflows, possibly lowering the costs and improving the resulting quality.

Against this background, within this thesis, a new and robust technique for camera pose estimation and rectification of uncalibrated stereo cameras based on a new method to estimate the fundamental matrix is proposed. The approach is subsequently enhanced towards trifocal setups involving a new estimation method for the trifocal tensor. To rectify the images acquired by uncalibrated cameras, a suitable feature detector is required. In this context, a new feature descriptor (SKB) is proposed and compared to existing descriptors such as SIFT, SURF or BRIEF. The different algorithms are combined, extended by new functions to calculate important stereoscopic parameters, and made accessible through an intuitive graphical user-interface which allows non-expert camera personnel to make use of it using an application which is called stereoscopic analyzer (STAN). Finally a new multi-camera disparity estimation workflow is proposed and applied to a multi-camera setup suitable for the generation of display agnostic 3D content.

¹ An extensive list of 3D movie releases which is maintained by Andrew Woods, Co-Chair of the annual Stereoscopic Displays and Applications (SD&A) conference, can be found on the following website [Woods14]: <http://www.3dmovielist.com/>

Kurzfassung

Hintergrund und Motivation für die in der vorliegenden Dissertation getätigten Forschungsarbeiten ist die Digitalisierung der Kino- bzw. Filmproduktion, die vollkommen neue Arbeitsabläufe auf Grundlage von neuen Bilderverarbeitungsalgorithmen erlaubt. Ein besonderes Augenmerk liegt dabei auf stereoskopischen 3D Produktionen und Multi-Kamera-Produktionen.

Während eine Vielzahl an 3D Produktionen in den letzten Jahren in die Kinos kam, so blieb das zugrundeliegende Prinzip der stereoskopischen Wiedergabe das gleiche wie von Wheatstones [Wheatstone38] und Brewsters [Brewster56] vorgestellt, d.h. dem Betrachter werden für das linke und rechte Auge zwei Bilder mit leicht unterschiedlichen Perspektiven bereitgestellt. Wenn das Betrachten der 3D-Inhalte allerdings zu unnatürlichen Sehbedingungen führt, z.B. durch zu große Parallaxe, kann dies zu Unwohlsein bei der 3D-Wahrnehmung führen [IJsselsteijn00]. Folglich müssen besondere 3D-Produktionsregeln, wie in [Mendiburu08] und [Knorr12] beschrieben, beachtet werden, um hochwertige 3D-Inhalte zu produzieren. Dies beinhaltet eine genaue Kalibrierung der Kameras mit konsistenten elektronischen und optischen Parametern. Ferner müssen Stereo-Basis und Konvergenzebene der 3D-Szene angepasst werden. Ohne Hilfsmittel oder Assistenzsysteme kann die Auswahl geeigneter stereoskopischer Parameter wie in [Lipton82] beschrieben ein sehr mühevoller Vorgang sein, der gut ausgebildetes Personal benötigt und die Gesamtkosten einer Produktion ansteigen lässt [Buchs11]. Mit der Einführung von digitalen Kinokameras wurde es möglich, 3D-Videoströme zu analysieren und ggf. elektronisch zu korrigieren mittels stereoskopischer Bildverarbeitungsprozessoren (engl. „stereoscopic image processors“) [Zilly10b, Sony]. Diese vereinfachen die oben genannten Aufgaben und erlauben eine kostengünstigere 3D-Produktion bei gesteigerter Qualität des produzierten Materials.

Vor diesem Hintergrund wird in der vorliegenden Dissertation ein neues Verfahren für die Schätzung der Kamerapose und Stereo-Rektifizierungsparameter basierend auf einem neuen Verfahren zur Schätzung der Fundamentalmatrix vorgestellt. Der Ansatz wird ferner auf trifokale Kamera-Systeme erweitert mithilfe eines neuen Verfahrens zur Schätzung des trifokalen Tensors. Ferner wird ein neuer Merkmalsdeskriptor (SKB), der für die Korrespondenzpunktanalyse von unkalibrierten Kameras eingesetzt werden kann, vorgestellt und mit bestehenden Verfahren wie SIFT, SURF und BRIEF verglichen. Die genannten neuen Verfahren werden kombiniert und um Funktionen zur Berechnung von stereoskopischen Parametern, sowie einer graphischen Benutzeroberfläche, erweitert. Das Stereoscopic Analyzer (STAN) genannte Assistenzsystem soll auch Nicht-Experten die Produktion von guten 3D Inhalten ermöglichen. Schließlich wird ein neues Verfahren zur Multi-Kamera-Disparitäts-Schätzung vorgestellt und auf einen Multi-Kamera-Aufbau zur Erstellung von tiefenbasierten 3D-Inhalten für verschiedene Endgeräte angewendet.

Danksagung

Zunächst möchte ich Herrn Prof. Dr.-Ing. Thomas Sikora danken. Er hat mir ermöglicht, an seinem Fachbereich an der TU Berlin zu promovieren und hat mir mit seinen wertvollen Hinweisen entscheidende Impulse für meine Arbeit geliefert.

Mein besonderer Dank gilt Prof. Dr.-Ing. Peter Eisert. Peter hat sich von Beginn an intensiv mit meiner Forschungsarbeit auseinandergesetzt. Unsere Diskussionen über Inhalt und Struktur der Arbeit waren sehr inspirierend, stets ergebnisorientiert und dabei immer freundschaftlich.

Ferner möchte ich allen Gutachtern für das Interesse an meiner Arbeit und die Bereitschaft, als Gutachter hierfür tätig zu werden, danken.

Meinen ehemaligen Vorgesetzten Peter Kauff und Dr.-Ing. Ralf Schäfer möchte ich danken, dass sie mir am Fraunhofer HHI die notwendigen Ressourcen für meine Arbeit zur Verfügung gestellt haben. Ohne den Freiraum, der mir dort gewährt wurde, hätte ich meine Forschung nicht durchführen können. Die von ihnen akquirierten Projekte PRIME, 3D4YOU und MUSCADE haben uns als Gruppe ermöglicht, an spannenden und zukunftsweisenden Forschungsfragen zu arbeiten. Peter möchte ich ganz besonders für seine Unterstützung und sein Vertrauen in mich danken. Er hat mich an das wissenschaftliche Arbeiten herangeführt und mir früh die Chance gegeben, aktiv eigene Forschung zu betreiben und meine Ideen in die Realität umzusetzen. Ohne ihn gäbe es den STAN nicht.

Mit meinen Kollegen Marcus Müller und Christian Riechert habe ich zahlreiche inhaltliche Diskussionen zu Tiefenschätzung, Multikamerageometrie und Bildverarbeitung geführt. Unsere Gespräche waren immer anregend, die gemeinsame Arbeit konstruktiv und zielorientiert. Durch gemeinsame Reisen zu Konferenzen in Europa, Japan und USA, durch gute Gespräche zum Feierabend in der Schnitzerei und nicht zuletzt durch das darauf folgende Feierabendbier wurden aus Kollegen Freunde.

Meinem aktuellen Vorgesetzten Dr.-Ing. Siegfried Föbel vom Fraunhofer IIS möchte ich danken, dass er mir ermöglicht hat, meine Arbeit in seiner Abteilung zu Ende zu führen. Dr.-Ing. Joachim Keinert danke ich für das gründliche Korrekturlesen meiner Arbeit und seine wertvollen Hinweise.

Danke an meine Frau Karin für ihre unermüdliche Geduld, Aufmunterung und Nachsicht in all den Jahren. Bei meinen Söhnen Viktor, Theodor und Arthur möchte ich mich entschuldigen für die vielen Urlaube, Wochenenden und Abende, an denen ich am Schreibtisch saß, anstatt Zeit mit ihnen zu verbringen. Jetzt ist "Papas Buch" fertig und das wird anders - versprochen!

Contents

1	Introduction	1
1.1	Motivation	1
1.1.1	Problem Statement and Relevance	1
1.1.2	Research Questions	2
1.1.3	Scope of the Thesis.....	3
1.2	State-of-the-Art 3D Production Workflow.....	4
1.2.1	Native Stereoscopic 3D Production with Two Cameras	4
1.2.2	2D to 3D Conversion.....	7
1.2.3	Depth-Based and Multi-Camera 3D Content Production.....	10
1.3	Dissertation-Overview.....	12
1.3.1	Publications related to the Thesis	12
1.3.2	Novelties and Contributions	14
1.3.3	Structure of the Dissertation.....	16
2	Theoretical Background	19
2.1	Geometry of Stereoscopic 3D Reproduction.....	19
2.1.1	Functional Concept of 3D Displays	19
2.1.2	Basic Geometric Concepts	22
2.1.3	Horizontal Image Translation and Parallax Range.....	24
2.2	The Human Visual System and Depth Perception	25
2.2.1	Binocular Depth Cues	26
2.2.2	Monocular Depth Cues.....	26
2.2.3	3D Perception Conflicts	27
2.3	Geometrical Implications for Stereoscopic 3D Production.....	30
2.3.1	Comfortable Viewing Range and Depth Budget.....	30
2.3.2	Geometrical Concepts of 3D Acquisition.....	31

2.3.3	Adjustment of Inter-Axial Camera Distance	32
2.3.4	Adjustment of the Convergence Distance and Horizontal Image Translation	33
2.3.5	Mechanical Alignment and Setup of the 3D Rig.....	33
2.4	Projective Geometry	33
2.4.1	Basic Camera Geometry	34
2.4.2	Two-Camera Geometry	38
2.4.3	Three-Camera Geometry	43
2.5	Feature Point Matching	46
2.5.1	Introductory Remarks	46
2.5.2	Interest Point Detection and Related Work	47
2.5.3	Interest Point Description and Related Work	52
2.5.4	Interest Point Matching	53
2.6	Disparity Estimation.....	55
2.6.1	Related Work.....	55
2.6.2	Stereo Disparity Estimation.....	56
2.6.3	Similarity Analysis	57
2.6.4	Left-Right Consistency Check	58
2.6.5	Disparity Map Filtering	59
2.7	Conclusion.....	60
3	Linearized Projective Geometry	63
3.1	Taylor Expansion for Projective Entities	63
3.2	Linearized Computation of the Fundamental matrix.....	64
3.2.1	Point Correspondences	65
3.2.2	Linearization Approach.....	65
3.2.3	Estimation of the Linearized Fundamental Matrix.....	67
3.2.4	Choice of the Fitting Parameters	68
3.2.5	Model Fitting with RANSAC.....	68
3.2.6	Singularity Constraint.....	70
3.2.7	Rectifying Homographies.....	70
3.2.8	Results	71

3.3	Linearized Computation of the Trifocal Tensor	75
3.3.1	Ideal Geometric Setup	76
3.3.2	Degrees of Freedom for the Linearized Geometric Setup	77
3.3.3	Solving Set of Linear Equations	78
3.3.4	Trifocal Rectification	79
3.3.5	Results	80
3.4	Conclusion	82
4	Fast Feature Point Description and Matching	85
4.1	Introductory Remarks	85
4.2	Basic Properties of the SKB-Descriptor	85
4.3	Defining the Support Region	86
4.3.1	Overlapping Kernel Set Evaluation Regions (Type A)	86
4.3.2	Uniform Kernel Set Evaluation Regions (Type B)	87
4.4	The Kernels	88
4.4.1	Filter Response Binarization	89
	Variant A: 256 Bits	90
	Variant B and C: 512 Bits	90
4.5	Matching Approach	90
4.5.1	Additional Matching Constraints	91
4.6	Evaluation of SKB	91
4.6.1	SKB vs. GLOH, SIFT and Cross Correlation	92
4.6.2	SKB vs. BRIEF	94
4.6.3	Application to Real-Time Stereo Matching	96
4.7	Conclusion	97
5	Assisted 3D Production	99
5.1	Introductory Remarks	99
5.2	Related Work	100
5.3	Overview of the Assisted 3D Production Setup	100
5.3.1	Stereo Rig with Mechanical Alignment Ability	101

5.3.2	Frame Grabber.....	101
5.3.3	Stereo Analysis Engine	101
5.3.4	Real-Time Rectification Engine.....	102
5.3.5	Post-Production Unit	102
5.3.6	Motor-Control Unit	102
5.3.7	Graphical User Interface.....	102
5.4	Algorithms for the Geometry Analysis	103
5.4.1	Robust Feature Detection	103
5.4.2	Temporally Filtered Pose Estimation	105
5.4.3	Temporally Consistent Stereo Image Rectification.....	106
5.5	Algorithms for the Depth Bracket Analysis	108
5.5.1	Calculation of a Disparity Histogram.....	109
5.5.2	Automatic Adjustment of the Horizontal Image Translation (HIT).....	111
5.5.3	Automatic Derivation of the Inter-axial Distance	113
5.6	Comparison with Legacy Workflows.....	114
5.6.1	Stereo Rig Calibration and Rectification.....	114
5.6.2	Dynamic Convergence Plane Adjustment.....	120
5.6.3	Adjustment of the Inter-axial Distance.....	123
5.7	Conclusion.....	126
6	Mixed Baseline Stereo Estimation	129
6.1	Introductory Remarks.....	129
6.2	Multi-Camera Content Acquisition and Pre-Processing	129
6.2.1	The MUSCADE Multi-Camera Setup.....	129
6.2.2	Calibration of the Linear Camera Array.....	131
6.2.3	Multi-Camera Rectification.....	131
6.3	Stratified Mixed-Baseline Disparity Estimation.....	132
6.3.1	Mixed Baseline Stereo Setup.....	132
6.3.2	Initial Disparity Estimation	134
6.3.3	Left-Right Consistency Check	134
6.3.4	Normalization of the Disparities	135

6.3.5	Merging of the Inner Disparity Maps	136
6.3.6	Filtering Inner Disparity Maps	138
6.3.7	DIBR of Disparity Maps from Inner to Outer Views	139
6.3.8	Merging of the Satellite Disparity Maps	140
6.3.9	Filtering the Satellite Disparity Maps.....	140
6.4	Results	141
6.4.1	Offline MVD4 Generation	141
6.4.2	Real-Time MVD4 Generation	145
6.5	Conclusion.....	145
7	Conclusion and Outlook.....	147
7.1	Summary	147
7.2	Main Contributions.....	148
7.2.1	Linearized Fundamental Matrix	149
7.2.2	Linearized Trifocal Tensor	149
7.2.3	Feature Point Descriptor SKB	150
7.2.4	Algorithms for the Simplified 3D Production	150
7.2.5	Mixed Baseline Multi-View Video plus Depth Generation	150
7.3	Discussion & Outlook	151
8	Appendix	153
8.1	Stereoscopic Test Productions.....	153
8.1.1	Cebit 2010	153
8.1.2	Berliner Philharmoniker	154
8.1.3	Fantastische Vier	155
8.1.4	Marina and the Diamonds	156
9	Glossary.....	157
9.1	Technical Terms	157
9.2	Abbreviations and Acronyms	157
9.3	Latin and Mathematical Symbols.....	158
10	Bibliography	159

10.1	Publications by the Author	159
10.2	Other Publications	160

1 Introduction

1.1 Motivation

1.1.1 Problem Statement and Relevance

The workflow for stereoscopic 3D productions for digital cinema and 3D live broadcast has increased in quality and efficiency in the past years. This is due to the fact that on one hand, a wide range of improved equipment such as 3D rigs is available, e.g. beam-splitter or side-by-side rigs in different sizes and configurations [Sony12]. On the other hand, there is today a better understanding of important 3D production rules [Mendiburu08, Zilly11b, Knorr12] which yield to a better overall production quality.

Nevertheless, the production costs, especially for live events, are still quite high, as additional and well trained personnel is required and additional equipment is needed compared to a standard 2D production. Thus, there is a clear demand for more cost efficient production workflows [Buchs11]. What adds to the challenge is that a high level of 3D quality needs to be maintained. While a poor overall 2D production quality might decrease the quality of experience, this is not an option for stereoscopic 3D [Knorr12]. In fact, a poor 3D production quality can result in a very bad user experience including headaches and eye-strain [IJsselsteijn00].

In this context, an important but very time consuming part of the current stereoscopic 3D workflows is the need to align the 3D rigs mechanically. Expert knowledge is required to adjust the two cameras precisely and to choose an appropriate inter-axial distance, or camera baseline as well as a suitable convergence plane. Any mechanical misalignment, such as unwanted roll or tilt errors, would result in an impaired stereoscopic image, which yields to headache or visual fatigue when watched. Methods to speed up the calibration and alignment process are therefore helpful for any stereoscopic 3D production facility [Sony12].

Another important development in the field of 3D-TV is the advent of auto-stereoscopic displays which enable a 3D sensation without glasses. For these displays, a high number of views are required, while the number of views differs from display to display [Dodgson05]. Moreover, this number of views will grow in the near future when 4k LCD panels are being used for auto-stereoscopic displays. In consequence, the required parallax for the different views exceeds the depth volume which is inherent to standard stereoscopic 3D setups. Thus, alternate camera configurations, involving multi-camera setups with e.g. four cameras are required [Smolic08, Zilly13]. An efficient way for the content generation using multi-camera setups is therefore required for a successful introduction of next generation 3D-TV services.

1.1.2 Research Questions

Against this background, and given the fact that the today's 3D content production chain is based on digital cinema and digital television techniques, the question is raised which algorithms from the field of digital image processing can facilitate the workflows and help to accomplish the goal of an efficient 3D production of stereoscopic 3D content. Moreover, how do these algorithms need to be extended towards multi-camera setups?

1.1.2.1 Image Rectification

Prospective candidates are techniques involving camera pose estimation and image rectification. In fact, huge progress has been achieved in the past years regarding self-calibration techniques involving e.g. the estimation of the epipolar geometry, camera pose, and similar geometric entities [Hartley04]. Closely linked to the two mentioned techniques is the image rectification technique, which allows for eliminating vertical disparities within a stereo pair. This property is suitable not only for computer vision algorithms such as disparity estimation, but is also valuable for improving the visual quality of stereoscopic 3D images. As one would want to apply an image rectification to already existing 3D content, the rectification technique should not rely on the use of calibration patterns. Moreover, important extrinsic and intrinsic camera parameters can change during a 3D live transmission (e.g. if zoom-lenses are used [Wu13]), and a pattern based calibration cannot be conducted during a live event. Consequently, image rectification techniques for stereo camera setups will play an important role within this thesis. Moreover, an extension towards a multi-camera setup is proposed and examined.

1.1.2.2 Feature Detection & Matching

Image rectification techniques suitable for uncalibrated camera systems rely on point correspondences between the different camera views which can be used to estimate the geometry between the cameras. Hence, an important research question is linked to the problem how reliable point correspondences can be established between the involved cameras, e.g. by using feature point detectors, descriptors, and feature point matching algorithms. Indeed, also in this research area, convincing algorithms such as SIFT [Lowe04] and SURF [Bay08] have been developed in the past years which are robust enough to find reliable point correspondences, even if the image regions are distorted due to parallax effects, non-matching focal lengths, focus mismatches, or similar effects. In fact, it is important to take these image impairment effects into account, as they often occur in a 3D production scenario involving two or more cameras. Consequently, a feature descriptor, denominated as SKB, which has been designed for the special needs of a 3D production environment is proposed and examined within this thesis. Compared to SIFT and SURF, the description and matching process is less complex and can thus be evaluated at a higher update rate while maintaining a comparable matching quality.

1.1.2.3 Assisted 3D Production Workflow

The workflow for the production of stereoscopic 3D content is more complex than the respective 2D production workflow as two camera views need to be generated and stereoscopic parameters such as the convergence plane and inter-axial distance need to be controlled [Mendiburu08]. Hence, the question is raised, which kind of assistance systems could facilitate the work of the stereographer² and camera personnel involved in a 3D production. How could the underlying algorithms help the stereographer and how should the interaction between the user and a processing unit look like? For instance, is it possible to automatically adjust the convergence plane, an important stereoscopic parameter which is traditionally controlled by a dedicated camera assistant called convergence puller. In the recent years, commercial stereo image processors and camera assistance systems became common tools at today's film sets, TV studios, and production facilities [Sony, 3ality, Binocle]. Against this background, a stereo assistance system comprising a set of new image processing algorithms based on [Zilly10b] is presented within this thesis. Beside the basic algorithms mentioned above such as image rectification and feature matching, advanced comfort functions are presented. These comprise a set of automated functions for image rectification and adjustment of the convergence plane and stereo baseline. The user can interact with the system using a graphical user interface GUI.

1.1.2.4 Multi-Camera Content Creation

The efficient production of multi-camera content is a key issue to enable a successful introduction of next generation 3D services using auto-stereoscopic displays and light-field displays [Zilly13]. These displays have in common, that a high number of views of the scene from different viewpoints are required [MüllerK08]. The question is raised, how this content can efficiently be produced, without for instance, involving as many cameras, as views required by the display. In fact, it is suitable to use a limited number of cameras and to generate the remaining views by means of Depth Image Based Rendering (DIBR) [Fehn04]. However, a set of camera images along with high quality disparity maps is required for DIBR algorithms [MüllerK11]. Within this thesis, a workflow is presented to generate Multi-view Video plus Depth content with four cameras (MVD4). The workflow, which is based on an existing disparity estimator designed for a two-camera setup, combines the stereo disparity maps to generate a set of disparity maps with high depth resolution and reliability.

1.1.3 Scope of the Thesis

Within this thesis, the mentioned research questions are tackled with the aim to improve the quality and cost efficiency of stereoscopic 3D productions and multi-camera productions.

² The task of the stereographer is to ensure that all relevant 3D production rules are met [Mendiburu08].

1.2 State-of-the-Art 3D Production Workflow

Different techniques exist for the production of 3D content. The method chosen depends on the target device, the type of content but also on budgetary decisions. For instance, the classic approach for a stereoscopic 3D production and 3D live productions in special, involves two cameras, mounted side-by-side or on a beam-splitter rig, capturing the content which will later be presented to the left and the right eye of the spectator [Lipton82, Lipton01, Mendiburu08]. Nevertheless, an alternate production workflow consists of capturing the content with one camera only, and creating the second required view using 2D-to-3D conversion algorithms, using semi-automatic approaches in post-production, e.g. taking advantage of a camera motion [Knorr06, Zhang07], or automatically using sophisticated algorithms [ZhangL11]. On the other hand, many today's 3D movies involve computer generated content. In this case, the whole 3D scene geometry is digitally available and can be controlled according to production guides such as [Lipton97]. The two stereo views are rendered using ray-tracing or similar techniques. Yet another production workflow is required to create content for multi-perspective displays, such as auto-stereoscopic and light-field displays typically involving a multi-camera configuration such as camera arrays [Matusik04, Wilburn05, Zitnick04]. Finally, dedicated depth sensors, such as time-of-flight cameras, or systems based on structured light exist to create depth maps. By combining the depth maps with existing texture information, virtual views can be generated by means of Depth Image Based Rendering (DIBR) [Fehn03a, Fehn03b]. Finally, to capture content for Free Viewpoint Video (FVV), a camera dome is required, i.e. a set of cameras capturing the same scene from different viewpoints [Smolic11b, Tanimoto06].

This thesis concentrates on content capturing using two or more cameras mounted on a stereo-rig or a linear camera array, thus a brief overview of current 3D production workflows will be given. The native generation of stereoscopic 3D content will be described in sub-section 1.2.1, followed by an overview of 2D to 3D conversion in sub-section 1.2.2. Section 1.2 closes with an overview of existing techniques for the depth-based content creation in sub-section 1.2.3.

1.2.1 Native Stereoscopic 3D Production with Two Cameras

The most obvious approach to produce stereoscopic 3D content is to capture the scene with two cameras. For this purpose, two different setups are widely used. In the so-called side-by-side configuration, the two cameras are placed next to each other, inducing a horizontal translation of the camera centers [Mendiburu08]. However, the minimal distance between the optical axes, denoted as B in Figure 1.1 (a) is given by the dimensions of the camera body and the diameter of the lenses. In consequence, small cameras and lenses are usually used if the side-by-side rig is placed near to the scene in order not to exceed a given depth budget [Mendiburu12]. In fact, a typical inter-axial distance for this scenario is 3-7 cm [Zilly11b]. An example for a side-by-side rig with small point-of-view cameras is given in Figure 1.2 (a). In contrast, a side-by-side rig with large cameras and lenses can be used without exceeding the depth budget when it is placed far away from the scene, e.g. when

shooting from a helicopter or the tribune of a sport stadium [Grau09]. A detailed introduction to 3D camera geometry and resulting implications for the inter-axial distance will be given in the sections 2.1 through 2.3.

In order to overcome the limitations of the inter-axial distance in a side-by-side configuration, beam-splitters, or mirror-rigs are widely used in stereoscopic 3D production. The concept is illustrated in Figure 1.1 (b). The optical axis of the first camera (here shown in blue) passes through a half-transparent mirror while the optical axis of the second camera (here shown in red) is reflected by the half-transparent mirror. The advantage of this concept is that the inter-axial distance can be freely adjusted and even set to zero, which can be useful for calibration purposes (cf. section 5.6.1). This allows using cine-quality cameras (with large camera bodies) and lenses mounted on a rig which is near the scene, or, in case of a Steadicam even within the scene. The flexibility regarding the inter-axial distance comes at a price of losing in theory 50% of the light received by each camera due to the half-transparent mirror. This corresponds to one f-stop. In practice, the amount of light which is transmitted and reflected is not equal, thus it happens that one camera receives significantly less than 50% of the luminance. Moreover, the half-transparent mirror affects the color temperature of the transmitted light [Mendiburu12]. Thus, a colorimetric correction of luminance and color temperature is usually required when using beam-splitters. Finally, polarized light emitted from water surfaces or metallic surfaces behaves differently under reflection and transmission. In consequence, content-aware image corrections might be required if large or important parts of the captured scene are affected by polarized light [Routier12]. An example of a beam-splitter rig with cine-quality cameras is given in Figure 1.2 (b).

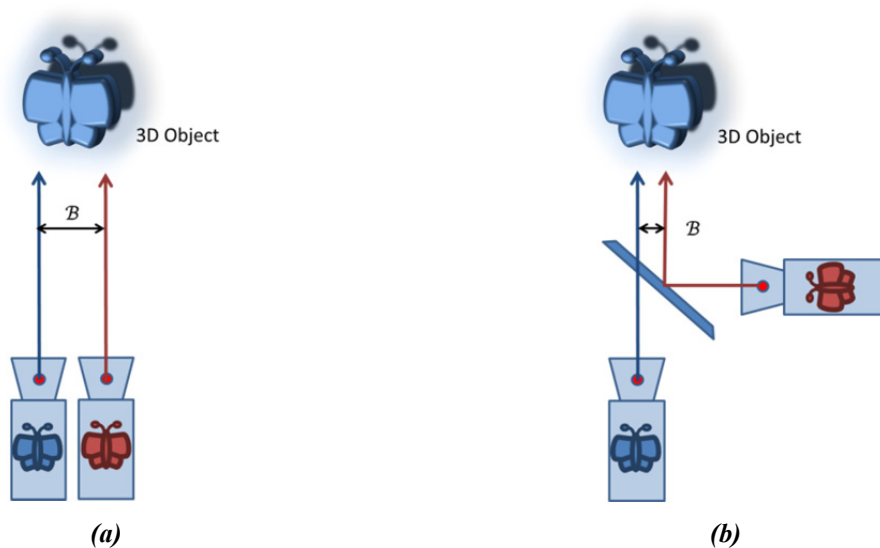


Figure 1.1. Illustration of the optical axes: (a) Side-by-Side configuration, (b) Beam-splitter configuration. The minimal stereo-baseline (also called inter-axial distance) B corresponds in the side-by-side configuration (a) to the size of the camera bodies and lenses, while in the beam-splitter configuration (b), smaller stereo-baselines can be realized.

In order to ensure that the optical axes point in the desired directions, a precise mechanical alignment of the cameras and the mirror (in case of a mirror-rig) is mandatory to produce good 3D content. Details on camera pose estimation and image correction techniques which can eliminate geometric distortions are discussed in chapter 5.



Figure 1.2. Picture of two common stereo-rig configurations: (a) In a side-by-side configuration, the minimal stereo baseline is defined by the width of the camera housings or bodies and the diameter of the lenses. (b) Using a beam-splitter rig, the stereo baseline or inter-axial distance B is freely adjustable.

1.2.1.1 Main Tasks involved in a 3D Production

The role of the stereographer is to ensure that all rules for a proper stereoscopic 3D production are respected. The 3D rig technician takes care of the mechanical alignment and the convergence puller of a suitable convergence plane. In this context, three important tasks can be identified:

- Mechanical calibration of the stereo rig;
- Adjustment of the inter-axial distance;
- Adjustment of the convergence plane.

Trained stereographers are able to perform these tasks by analyzing the images from the left and right cameras. Assuming that the images are displayed in anaglyph mode or by showing the difference of the luminance signals, it is possible to measure the amount of vertical and horizontal disparities. If, for instance, one of the cameras is rotated around its optical axis, a characteristic pattern of vertical and horizontal disparities is the result. The stereographer can then adjust the mechanical alignment of the stereo rig accordingly. Details will be given in section 5.6.1.

To adjust the inter-axial distance properly, the amount of parallax within the scene needs to be measured. Again, trained stereographers can perform this task by the use of special grid lines, overlaid on the left and right views or by physically measuring the distances of the nearest object in the scene and the scene background and calculating the optimal inter-axial distance using dedicated formulae which will be derived in section 2.3. However, this task can be very time consuming and expert

knowledge is required. Additional details on current techniques for adjusting the inter-axial distance will be given in section 5.6.3.

Finally, the adjustment of the convergence plane needs to be performed by the convergence puller. Related concepts such as the comfortable viewing range and horizontal image translation will be explained in sections 2.1 to 2.3. Details on current techniques for the adjustment of the convergence plane will be described in section 5.6.2.

1.2.1.2 Stereoscopic Image Processors

In order to simplify the main tasks of the stereographer, a couple of related assistance systems have been developed during the last years. Well established systems are the Stereoscopic Image Processor (SIP) from 3ality Digital [3ality], the Multi Image Processor (MPE 200) from Sony [Sony], the Disparity Tagger and Disparity Killer from Binocle [Binocle] and the stereoscopic analyser (STAN) from Fraunhofer HHI [Zilly10b]. The latter system is used as reference system for this thesis. It employs different algorithms invented and developed in the context of this thesis. A close look to the system will be given in chapter 5.

1.2.1.3 Stereoscopic 3D Post Production

The post-production for stereoscopic 3D content contains a set of additional steps compared to a standard 2D post-production. Remaining inconsistencies between the left and right camera regarding colour temperature and geometric distortions need to be corrected. There exists a multitude of 3D post-production tools on the market. A good overview of the different tools is given by [Mendiburu12].

1.2.2 2D to 3D Conversion

An alternative workflow for the stereoscopic 3D production is the 2D to 3D conversion which can be done in a semi-automatic process [Knorr06] or automatically using algorithms proposed by [ZhangL11]. Within this workflow, the scene is captured by a single camera only. Consequently at least one additional view needs to be generated artificially. There are different motivations to perform a 2D to 3D conversion and a good overview of related techniques is given in [Smolic11a]. First of all, there is a lot of legacy 2D content, which was not shot using two cameras and hence must be converted in order to reproduce it in 3D. Although the amount of available 3D content has increased considerably in the past years [Woods14], the majority of film material is still in 2D. Hence, to overcome a potential shortage on 3D content, the 2D to 3D conversion can be a suitable way.

On the other hand, there are feature film productions which target a 3D release but are nevertheless shot using a single camera, relying on a suitable quality of the conversion process. The quality of the artificial views highly depends on the type of content to be converted as well as the amount of human interaction, thus inducing budgetary and timing decisions. For instance, there exists broadcast

equipment which offers a built-in real-time and fully automatic 2D to 3D conversion [Sony]. It is obvious that the quality of this conversion cannot compete by no means with the manual 2D to 3D conversion as performed for the Hollywood movie *Titanic* costing \$18 million according to [Giardina12]. Although these two examples differ considerably in the amount of effort spent for the conversion, the basic techniques and challenges remain the same. Three main steps have to be conducted: depth map generation, pixel shifting, and occlusion filling.

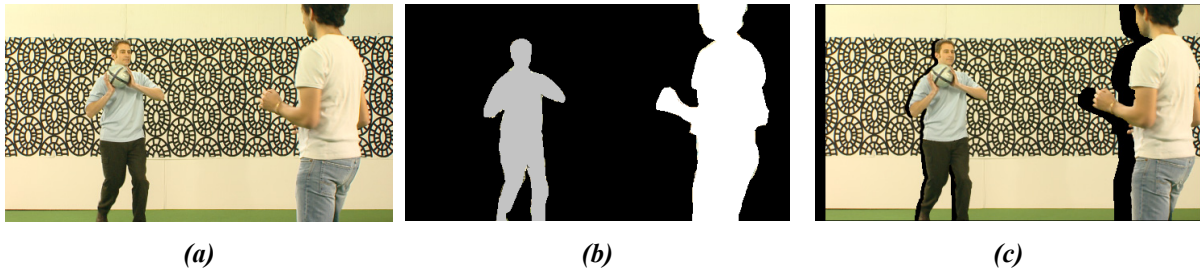


Figure 1.3. (a) *Original 2D View.* (b) *Manually generated Depth map.* (c) *Objects shifted horizontally according to their depth value assigned in a depth map³.*

1.2.2.1 Manual Depth Map Generation

In a first step, a depth map needs to be generated. This can be done manually in post-production using rotoscoping where a manual segmentation of objects is performed. This is a time consuming task, as this process needs to be performed for each individual frame of a video sequence or movie although specialized tracking software such as Mocha Pro 4.1 [Imagineer] or Silhouette fx v5 [SilhouetteFX] can support the artist. Subsequently, a depth value is assigned to each segmented object. The exact depth values are usually not required, but the ordering of the objects in depth space, or z-direction, needs to be correct. For a human operator, this is usually not very challenging, as monocular depth cues can be analyzed and interpreted (cf. section 2.2.2). An example is given in Figure 1.3 where a depth map (b) needs to be generated for the 2D image (a). The human observer easily recognizes, that the person wearing the white T-shirt in the right part of the image is nearer to the camera than the person with the blue pull-over which is about to throw the ball. Consequently, the person nearer to the camera is assigned another depth value (here light gray) than the person farther away from the camera (here dark gray). The background is even farther away and consequently, an even darker value is assigned to it (here black). Details about monoscopic and binocular depth cues will be given in section 2.2. In the case of Figure 1.3, the *relative size* of the two persons is the strongest indicator for their relative depth values.

Other techniques for the automated generation are so-called surrogate depth maps from 2D images try to take advantage of monocular depth cues, such correlations between chrominance and depth value under certain lighting conditions or by assuming that a given scene matches a template structure such e.g. *outdoor scene* or *tunnel structure*.

³ The images from Figure 1.3 (a) and (c) have been created in the context of the 3D4YOU project [3D4YOU] and have previously been published in [Zilly13]. Figure 1.3 (b) is only a sketch of a depth map which has not been used to create Figure 1.3 (c).

1.2.2.2 Pixel Shifting

Once the depth values are available, a new view can be generated by shifting each object horizontally according to its respective depth value. The result is shown in Figure 1.3 (c). The person near to the camera has shifted more than the person which stands farther away from the camera. This effect correlates to the horizontal disparities which will be discussed in more detail in section 2.6. The effect is similar to what one can observe when sitting in a train and looking through a window: Near objects pass quickly, while the farther objects pass slower. It is also possible to apply a non-linear shifting to objects in the scene to create depth effects which might be less realistic from the pure geometrical point of view but still offering a better 3D sensation as better advantage of a given depth budget can be taken [Holliman04, Cheng08, Lang10].

Pixel shifting is the basic principle behind Depth Image Based Rendering (DIBR) which can be used to synthesize new views using depth maps and existing views. Basic DIBR techniques will be applied in chapter 6.

1.2.2.3 Occlusion Filling

Once the pixel shifting has been performed, as shown in Figure 1.3 (c), the areas which are shown in black need to be filled. These regions were originally occluded in Figure 1.3 (a), and now became exposed.

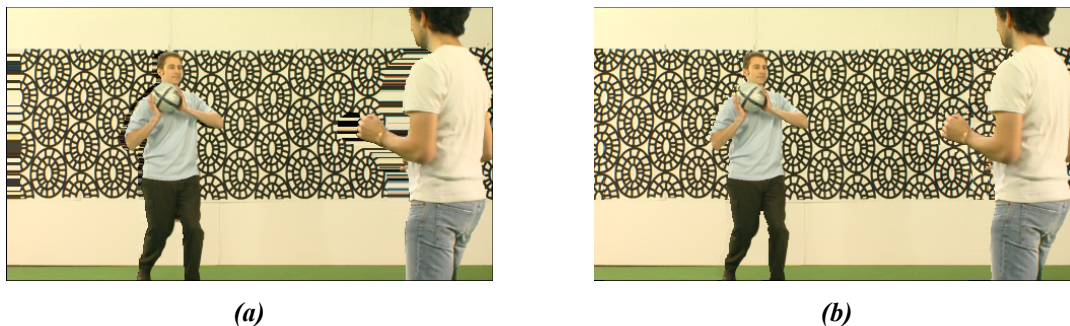


Figure 1.4. (a) Naive inpainting (b) Inpainting using patch matching⁴.

In Figure 1.4 two different methods are applied to fill the disoccluded image areas with texture. The simplest way to fill the holes is to perform a pixel repetition of the background pixels. This leads to acceptable results if the background has a homogeneous texture (e.g. the green and white backgrounds in Figure 1.4) but severe artifacts become visible when the background has a complex structure (e.g. the black circular patterns). In the latter case, more sophisticated inpainting techniques such as patch matching [Köppel10, Ndjiki-Nya10] are required. The use of the mentioned occlusion filling techniques is not limited to the case of 2D to 3D conversions but also applies to virtual view synthesis in general.

⁴ The images from Figure 1.4 have been created in the context of the 3D4YOU project [3D4YOU] and have previously been published in [Zilly13].

1.2.3 Depth-Based and Multi-Camera 3D Content Production

The generation of display agnostic 3D-TV content is a field of research since many years in computer vision and several approaches have been investigated in the past ranging from multi-camera systems to depth range sensors, structured light techniques, or a combination of these techniques. In fact, the approaches have in common that a set of real or virtual camera views needs to be reproduced at the display. The concept is visualized in Figure 1.5 using the example of the stereo-to-multi-view conversion.

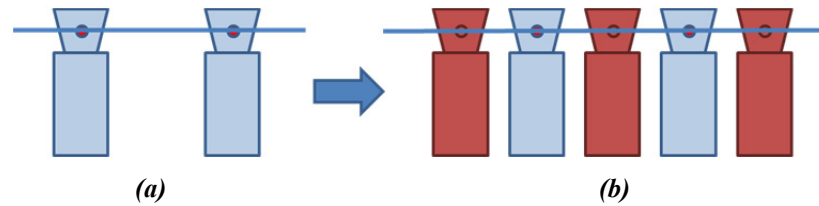


Figure 1.5. Concept of the stereo-to-multi-view conversion. Given a stereo pair of original cameras views (a), a set of interpolated and extrapolated views (red) are generated (b). A typical approach is to perform a disparity estimation with subsequent depth-image-based-rendering. The blue cameras in (b) are the two original views from (a) which were reused.

Given an original stereo pair (a), a set of interpolated and extrapolated views are generated (b). Different approaches for the generation of multi-view data are presented in the following.

A typical approach is to estimate disparity maps prior to apply depth-image-based-rendering as proposed in [Riechert12]. The same approach is suitable to generate a virtual stereo pair, e.g. to virtually adapt the stereo baseline or to rearrange objects in 3D space in a non-linear way to improve the 3D sensation of stereoscopic content as proposed by [Lang10]. The quality of the synthesized views usually decreases with the distance to the nearest original view. Moreover, interpolating views between original views yields better results than view extrapolation [Zilly2013].

A typical multi-camera setup is shown in Figure 1.6. A set of 16 machine vision cameras was used to capture 3D test material [Feldmann08]. Hitachi 3-chip CCD progressive scan RGB cameras (HV-F31CL-S1) with a resolution of 1024x768 pixels were used within the setup.



Figure 1.6. Camera-Array with 16 machine vision cameras used to capture the test sequence Book Arrival [Feldmann08].

A sample frame of the rectified video sequence is shown in Figure 1.7. The need for a very precise mechanical alignment of the linear camera-array made the content creation very time consuming. To keep the baseline of this multiple side-by-side setup small and to reduce the overall costs of the system, only machine vision cameras and no cine-quality cameras were used within the setup of [Feldmann08]. Hence, the visual quality might not fit all requirements of the cinema production.



Figure 1.7. Sample frame for the multi-camera sequence *Book Arrival* which was shot using 16 cameras [Feldmann08].

An alternative way for creating virtual views is the image domain warping (IDW). In [Stefanoski13], the authors describe the generation of a warp field based on robust feature point matches which are generated using the descriptor SKB (presented in [Zilly11c] and chapter 4). An approach called hybrid 3D using three cameras on a common baseline, with a high quality cinema camera in the center and two lower quality satellite cameras was recently proposed by Tanger et al. [Tanger13].

Beside the depth estimation and view rendering itself, complete 3D-TV chain requires techniques for data acquisition, coding, and transmission which includes the research question for an appropriate representation format for such content [Zilly13]. Different research projects investigated the above-mentioned research question. A video-plus-depth format was investigated within the European research project ATTEST [Redert02, Fehn02, Fehn04]. As acquisition device, a ZCam [Iddan01] which is able to generate a depth map was used. An approach based on disparity estimation for tele-presence applications was investigated within the European research project 3D-Presence [Schreer08, Feldmann09a, Divorra10]. Within the European research project 3D4YOU [Bartczak11, 3D4YOU], a setup with four cameras, with a narrow baseline involving two cameras mounted behind a beam-splitter and a wide baseline composed of two satellite cameras was used. The satellite cameras were not mounted on a common baseline with the cameras inside the mirror box. Within the European research project MUSCADE [Muscade], a similar specialized multi-camera rig with four cameras and an associated production workflow was developed. A linear camera array was used with two cameras mounted on a standard beam-splitter rig and two satellite cameras outside the mirror box. In contrast to the approach followed in the 3D4YOU project, all four cameras were mounted on a common baseline. A multi-camera disparity estimation workflow based on the MUSCADE setup will be presented in chapter 6.

An overview of different depth-based content acquisition setups can be found in [Zilly11b] while a summary of different research activities in the field of multi-camera and depth-based content production is given in [Ho10], [Grau11], and [Smolic11b]. An overview of research activities for free-viewpoint television (FVV) is given in [Tanimoto06].

1.3 Dissertation-Overview

1.3.1 Publications related to the Thesis

In the context of this thesis, a set of papers at different peer-reviewed international conferences, articles in peer-reviewed international journals as well as a book chapter were published. In this section, an overview of the most relevant publications is given and explained how they relate to the novelties and contributions described in more detail in the subsequent section 1.3.2.

The book chapter [Zilly13] was published in the following book:

- *3D-TV System with Depth-Image-Based Rendering*, Springer New York. Ce Zhu, Yin Zhao, Lu Yu, Masayuki Tanimoto (Editors), Jan. 2013:

Within the book chapter [Zilly13], a strategy for display agnostic content creation for 3D displays is presented which was investigated in the context of the European Research project MUSCADE [Muscade]. Details on a trifocal multi-camera rectification algorithm and techniques for disparity estimation and view synthesis are given. The publication relates mainly to the contribution described in the sub-section 1.3.2.5.

Research articles were published in the following journals:

- Proceedings of the IEEE (PIEEE), Special Issue on *3D Media and Displays*, vol. 99, issue 4, 2011, [Zilly11b]:

In the article [Zilly11b], production rules required for a successful 3D content production are presented along with details on the human visual system and its capacity to perceive depth based on so-called monoscopic and binocular depth cues. A mathematical foundation of 3D capture, display and image-based depth analysis is given. A concept for a 3D assistance system is presented. The publication relates mainly to the contributions described in the sub-sections 1.3.2.1 and 1.3.2.4.

- Journal of Visual Communication and Image Representation, Special Issue on *3D Video Processing*, vol. 25, issue 4, 2014, [Zilly14]:

The article presents a system suitable for the real-time estimation of disparity maps using a linear camera array with mixed wide and narrow baseline. The multi-camera setup, multi-camera rectification and disparity estimation are described in detail along with results from generated using depth-image-based rendering techniques. The publication relates to the contributions described in the sub-sections 1.3.2.2 and 1.3.2.5.

Research papers were published in the proceedings of the following international conferences:

- 17th IEEE International Conference on Image Processing (ICIP), Hong Kong, Sept. 2010, [Zilly10b]:
In the conference paper [Zilly10b], an image-based system for the assisted stereoscopic 3D production is presented. The publication relates to the contribution described in the sub-section 1.3.2.4.
- 21st International Conference on Pattern Recognition (ICPR), Tsukuba, Japan, Nov. 2011, [Zilly12c]:
In [Zilly12c], an algorithm for the estimation of rectifying homographies for trifocal camera setups is presented which involves the estimation of a linearized trifocal tensor. The publication relates to the contribution described in the sub-section 1.3.2.2.
- 8th European Conference on Visual Media Production (CVMP), London, UK, Nov. 2011, [Zilly11c]:
In the conference paper [Zilly11c], a new descriptor for image features is proposed which is optimized for stereoscopic camera setups. The publication relates to the contribution described in the sub-section 1.3.2.3.
- 5th International Symposium 3D Data Processing, Visualization and Transmission (3DPVT'10), Paris, France, May 2010, [Zilly10a]:
In the paper [Zilly10a], an algorithm for the estimation of the fundamental matrix which is near the rectified state is presented. The publication relates to the contribution described in the sub-section 1.3.2.1.
- The True Vision – Capture, Transmission and Display of 3D Video (3DTV-CON), Zurich, Switzerland, Oct. 2012, [Zilly12b]:
In the paper [Zilly12b], an algorithm for the stratified disparity estimation using a mixed narrow and wide baseline camera setup is presented. The publication relates to the contribution described in the sub-section 1.3.2.5.
- 14th ITG Conference on Electronic Media Technology (CEMT), Dortmund, Germany, March 2011, [Zilly11a].
In the paper [Zilly11a], an overview of field-tests of the assistance system presented in [Zilly10b] is given. The publication relates to the contribution described in the sub-section 1.3.2.4.

A complete list of authored or co-authored publications related to the topic of this thesis is given in the bibliography in section 10.1.

1.3.2 Novelties and Contributions

The novelties and contributions in this dissertation are twofold: a detailed overview of the state-of-the-art, theoretical foundations and related work presented in chapters 1 and 2 on one hand, and algorithmic contributions presented in chapters 3 to 6 on the other hand.

An overview of the field of stereoscopic 3D production is given, along with an overview of the relation between different multi-camera and depth-based image processing techniques and state-of-the-art 3D content production is given in chapter 1. In chapter 2, theoretical foundations of the 3D production process based on insights about the human visual system will be presented. This includes details of the 3D reproduction geometry and basic stereoscopic concepts such as inter-axial distance and convergence plane. Subsequently, concepts of underlying image processing techniques such as projective geometry, stereo- and multi-rectification, feature detection, feature description and matching, as well as disparity estimation will be explained.

In the sections 1.3.2.1 to 1.3.2.5, an overview of the algorithmic novelties and contributions is given.

1.3.2.1 Linearized Fundamental Matrix

When producing content for 3D cinema or 3D-TV, one goal is a perfectly aligned pair of stereo sequences. In fact, any misalignment of the cameras leads to vertical disparities. These vertical disparities in stereo pairs lead to eye strain and visual fatigue [Woods93]. Every stereo rig contains parts of finite mechanical accuracy. Moreover, thermal dilation changes the extrinsic parameters. When changing the lens' focus, the internal parameters such as the focal length can be affected [Fraser06]. In addition, lenses are changed during shootings, and the setup time is limited. When using zoom lenses, the focal length is changed over a wide range of values. Simultaneously, the principal point can shift due to the high number of lenses which are not fully concentric [Fraser06]. The motors for zoom level and focus do not synchronize exactly in the general case, so that slightly different focal lengths will occur [Fobker11]. Finally, these motors suffer from backlash which can be thought as a hysteresis curve which affects the zoom level. In consequence, it would be difficult to pre-calibrate a complete stereo rig and to generate meta-data in advance if all possible degrees of freedom such as backlash, thermal dilation, zoom lenses, changing focus and possible displacements of the half-transparent mirror due to shock and vibration of the mirror rig need to be taken into account. Therefore, a rectification algorithm is needed which performs reliably and which uses only point correspondences which can directly be extracted from the stereo image pairs. Thereby, the resulting rectified image pair should be suitable for watching, i.e. the rectification method needs to minimize any possible distortion. In addition, the convergence plane should not be changed because this is a critical stereo parameter for the 3D sensation. As a result, the rectification should not be done with respect to the plane at infinity [Fusiello08] but to a scene dependent plane. Against this background, a new rectification method is proposed in chapter 3.

1.3.2.2 Linearized Trifocal Tensor

Multi-camera systems such as linear camera arrays are commonly used to capture content for multi-baseline stereo estimation, view generation for auto-stereoscopic displays, or similar tasks. However, even after a careful mechanical alignment, residual vertical disparities and horizontal disparity offsets impair further processing steps. In consequence, the multi-camera content needs to be rectified on a common baseline. The trifocal tensor represents the geometry between three cameras and hence is a helpful tool to calibrate a multi-camera system, and to derive rectifying homographies. Against this background, in section 3.3 a new method for a robust estimation of the trifocal tensor specialized for linear camera arrays and subsequent rectifying homography computation based on feature point triplets is proposed. It is assumed that the camera geometry of the setup was designed for DIBR applications. In consequence, it is assumed that the geometric configuration is not far from the rectified state and that consequently a linearization is possible. The algorithm achieves vertical and horizontal alignment, i.e. horizontal disparities are proportional to each other after rectification. The ratio of the camera baselines can be extracted. The algorithm is suitable for uncalibrated cameras, as it uses feature point triplets and is robust against noise and outliers. Details of the proposed estimation method for the trifocal tensor will be presented in chapter 3.

1.3.2.3 Feature Point Descriptor SKB

State-of-the-art feature detectors distinguish interest point detection and description. The former is commonly performed in scale space, i.e. using a set of different image resolutions, while the latter is used to describe a normalized support region using histograms of gradients or similar derivatives of the grayscale image patch. This approach has proven to be very successful. However, the descriptors are usually of high dimensionality in order to achieve a high descriptiveness.

Against this background, a binarized descriptor which has a low memory usage and good matching performance is proposed. The descriptor is composed of binarized responses resulting from a set of folding operations applied to the normalized support region. A main property of the SKB is a low computational load and complexity. Its fast run-time enables near real-time updates of stereo rectification parameters. Details of the feature descriptor SKB will be presented in chapter 4.

1.3.2.4 Algorithms for the Simplified 3D Production

A set of new algorithms for the temporal consistent estimation of the 3D camera geometry, such as the generation of a disparity histogram for the derivation of the near and far clipping plane, is presented in chapter 5. The algorithms combine and make use of the rectification algorithm from chapter 3 and the feature descriptor SKB from chapter 4. In combination with a PC system with graphical user-interface, the algorithms are the core of a camera assistance system which supports the stereographer using comfort functions such as the automatic derivation of the convergence plane and the inter-axial distance.

1.3.2.5 Mixed Baseline Multi-View Video plus Depth Generation

Within the content production for stereoscopic 3D-TV displays, two different views need to be generated. The content is usually shot using two cameras as the glasses-based target devices require two views as input. Beside stereoscopic 3D-TVs, huge progress has also been achieved in the improvement of the image quality of glasses-free auto-stereoscopic displays and light-field displays. Concerning the latter two display families, the content production workflow more complex, as the number of required views not only differs considerably but is also likely to increase in the near future. A depth-based content creation workflow using high quality HD-TV cameras could be suitable to generate an arbitrary number of views for the different displays. Against this background, a new algorithm for the multi-camera disparity estimation using mixed stereo baselines is presented which is suitable for real-time execution. The setup is based on a four camera rig involving a central narrow baseline, with two cameras mounted on a standard beam-splitter rig known from stereoscopic 3D productions, and a wide baseline comprising of two satellite cameras mounted outside the mirror box. As all four cameras are positioned on a common baseline they form a linear camera array. In chapter 6, the multi-view video plus depth generation workflow optimized for the four-camera setup is described in detail.

1.3.3 Structure of the Dissertation

The remainder of the dissertation is structured as follows: In chapter 2, the theoretical background for the subsequent chapters is presented. This includes details about the geometry of stereoscopic 3D content reproduction (section 2.1), the human visual system (HVS) and depth perception in section 2.2, and resulting production rules for the acquisition of 3D content which have to be respected by stereographers (section 2.3). In section 2.4, fundamental concepts of the projective geometry are described along with basic concepts of the stereo and multi-camera rectification. In section 2.5, basic concepts of feature detection are presented before completing the chapter with section 2.6 where fundamental concepts of disparity estimation techniques are presented. In section 2.7 the chapter is concluded.

Chapter 3 starts with a description of how the concept of the Taylor expansion is applied to projective entities such as the rotation matrix or projection matrix (section 3.1). Subsequently, this concept is applied to the linearized estimation of the fundamental matrix (section 3.2) and the estimation of the trifocal tensor (section 3.3). The linearized projective entities are then used to derive rectification algorithms specialized for the case of nearly rectified stereo cameras (section 3.2) and linear cameras arrays (section 3.3). A quantitative evaluation of the performance of the proposed algorithms is carried out before concluding the chapter in section 3.4.

A new feature descriptor denominated as *semantic kernels binarized* (SKB) is presented in chapter 4. The chapter starts by introducing the basic properties of the descriptor (section 4.2) before describing

in more detail the several steps of the description process, i.e. the definition of the support region, the sampling of the support region (section 4.3) and the folding with a set of binary kernels (section 4.4). Matching strategies which take advantage of the binary feature vector of the descriptor are presented in section 4.5. A comparison with state-of-the-art feature point descriptors is performed in section 4.6 along with a detailed quantitative evaluation before concluding the chapter with section 4.7.

In chapter 5, algorithms for the temporal consistent analysis of stereoscopic 3D sequences are presented. An overview of the involved components is given in section 5.2. Based on the approaches for the linearized estimation of the fundamental matrix from chapter 3 and the feature descriptor SKB from chapter 4, algorithms for a temporal consistent camera pose estimation, rectification (section 5.3) and disparity histogram analysis are presented. The latter is subsequently used to derive the near and the far clipping plane of a scene in order to derive optimal stereoscopic settings for the convergence plane and the inter-axial distance (section 5.4). The proposed algorithms shall enable a simplified 3D production workflow. Consequently, a detailed comparison of the updated production workflow with legacy production workflows is performed in section 5.5 before concluding the chapter with section 5.6.

In chapter 6, a new workflow for the mixed baseline disparity estimation is presented. In section 6.2, the background of the multi-camera setup and the technical requirement are introduced. As an important preprocessing step, a multi-camera rectification based on the trifocal tensor estimated using the approach from chapter 3, and feature points matched using the approach from chapter 4, is applied. For a proper alignment of the multi-camera rig the assistance system from chapter 5 is used with modifications which take the multi-camera geometry into account. In section 6.3, the stratified approach for the multi-camera disparity estimation is presented. The approach combines disparity estimation from a narrow baseline which usually outputs dense disparity maps but with lower depth resolution on one hand, and a wide baseline estimation which usually outputs sparse disparity maps with high depth accuracy on the other hand. Results of the approach will be shown in section 6.4 before concluding the chapter with section 6.5.

In chapter 7, the main outcomes and findings of the thesis are summarized, discussed and concluded. Finally an outlook on future possible research is given.

2 Theoretical Background

In this chapter, the theoretical background for the subsequent chapters is presented according to the structure presented in section 1.3.3. The topics of the foundations are thereby closely related to the research questions presented in section 1.1.2, i.e. image rectification, feature detection and matching, assisted 3D production and multi-camera content creation.

The chapter starts with details about the geometry of stereoscopic 3D reproduction (section 2.1) before presenting details of the human visual system (HVS) and depth perception in section 2.2. Resulting production rules for the acquisition of 3D content which have to be respected by stereographers are presented in section 2.3. The topics discussed in the three sub-sections are the basis for the thesis in general and the stereoscopic assistance system presented in chapter 5 in particular. In section 2.4, fundamental concepts of the projective geometry are described along with basic concepts of the stereo and multi-camera rectification. These concepts constitute the theoretical background for chapter 3. In section 2.5, basic concepts of feature detection are presented which are important for the feature descriptor SKB presented in chapter 4. Fundamental concepts of disparity estimation techniques are presented in section 2.6. These techniques are used in chapter 6, where a multi-camera disparity estimation algorithm is proposed. Finally, the chapter is concluded in section 2.7.

2.1 Geometry of Stereoscopic 3D Reproduction

2.1.1 Functional Concept of 3D Displays

An overview of different 3D displays devices is given as content creation is usually performed with respect to a target device⁵. Against this background, the functional concept of common 3D displays is explained in the following.

2.1.1.1 Functional Concept of Stereoscopic Displays

A variety of glasses-based stereoscopic 3D displays exist on the market. Most of today's 3D displays are based on a Full HD panel, i.e. a total resolution of 2 Mega-Pixel is available [Mendiburu12]. Depending on the 3D technology, 1 or 2 Mega-Pixels are available per view. The former holds true for conventional line-polarized displays while the latter applies to shutter displays. In the case of cinema projectors, there exist single projector solutions as well as concepts involving two projectors. The basic principles are similar to 3D displays, i.e. a channel separation between left and right view has to be performed. An excellent overview of 3D cinema projector technologies is given in [Lipton01].

⁵ Parts of the content in this section have been previously published in [Zilly11b] and [Zilly13].

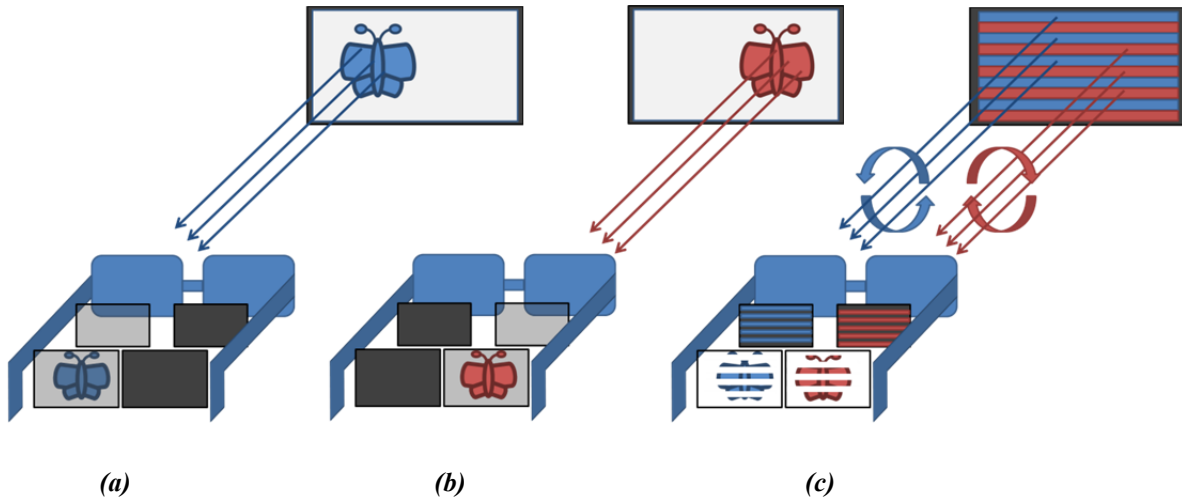


Figure 2.1. *Shutter glasses (a, b) or micro-polarizers (c) are used to establish a channel separation between the pixels dedicated to be seen by the left and right eye.*

Shutter displays show the image corresponding to the left and right view in a time sequential way. The observer needs to wear glasses with the ability to change the transparency of the left and right glasses according to the image presented in the display. In Figure 2.1 (a, b) this process is illustrated. In a first moment, the display shows the image dedicated to the left eye. The left glass is transparent while the right glass is made intransparent as shown in the Figure 2.1 (a). In the next moment, the display presents the image dedicated to the right eye, while the transparency of the left and the right glasses are toggled as shown in Figure 2.1 (b). To avoid cross-talk between the left and the right view, the point in time at which the left and right views are switched needs to be synchronized between the glasses and the display. Shutter glasses contain active visual components [Foessel09] hence they require batteries which increase their weight.

Another concept used to create the channel separation between the left and the right view is polarization. Therefore, micro-polarizers are placed in front of the 3D display. As shown in Figure 2.1 (c), in front of every odd line, a micro-polarizer with counter-clock wise polarization characteristic is attached while in front of every even line, a micro-polarizer with clock-wise polarization characteristic is used. Corresponding polarizing foils are attached on the left and respectively right glass worn by the user. As a result, the full horizontal resolution but only half of the vertical resolution of the displays is available for each view. Channel separation using polarized light can also be performed using 3D projectors [Foessel09]. In this case, two projectors are used with micro-polarizers in front of the optical system. The screen surface needs to preserve the polarization in order to avoid significant cross talk between the channels. A metallic surface can be applied to a screen to fulfill this requirement. Details of the screen design along with a discussion on new challenges arising from viewpoint dependent reflection properties of these screens can be found in [Lipton01].

Finally, one can achieve the required channel separation by inserting a color bias between the left and the right view, e.g. using red and cyan color filters as used for anaglyph glasses. The concept of

channel separation based on color filters, although in a much more sophisticated way, is used in projectors and glasses by Infitec⁶.

2.1.1.2 Functional Concept of Auto-Stereoscopic Displays

An illustration of the functional concept of auto-stereoscopic displays is given in Figure 2.2. By using lenticular lenses as shown in Figure 2.2 (a) or parallax barriers (b), the light rays emitted by pixels or sub-pixels corresponding to a certain view are focused to a sweet spot in front of the display [Lipton82].

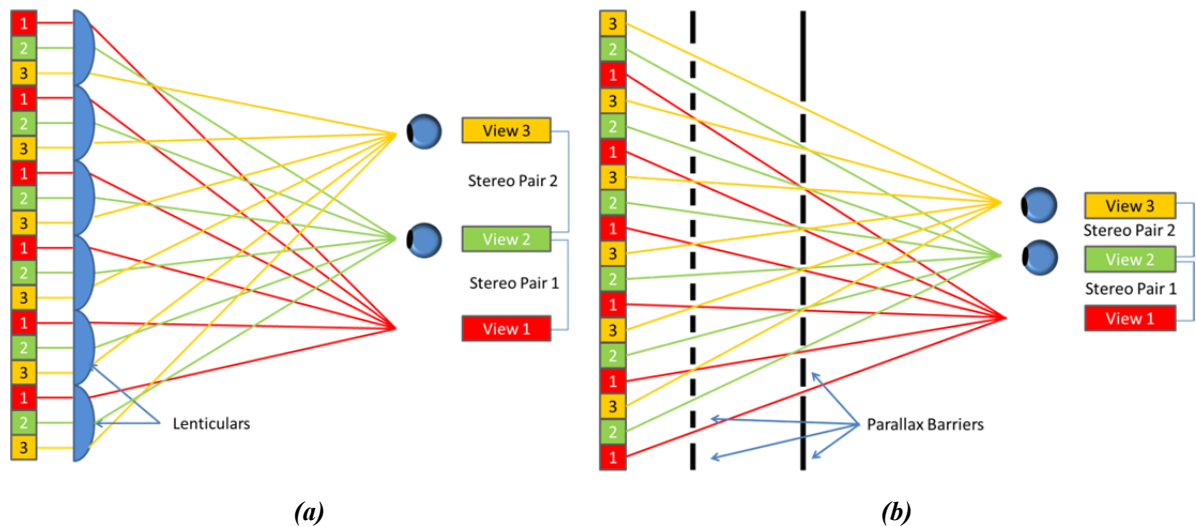


Figure 2.2. Function of lenticular lenses (a) and parallax barriers (b) used for auto-stereoscopic 3D displays.

Ideally, the distance between neighboring sweet spots corresponds to the inter-ocular eye distance t_{eye} . This allows the spectator to watch a valid stereo pair. In Figure 2.2 (a) and (b), three sweet spots corresponding to the views 1-3 are illustrated. In the drawing, the spectator currently watches the stereo pair 2, created by views 2 and 3. When moving the head to the left by an amount similar to the distance between two sweet spots, the other stereo pair 1 would be perceived, creating the illusion of head motion parallax. Ideally, the spectator is now able to see parts of the image which were occluded in stereo pair 2.

An overview of the technical principles of auto-stereoscopic displays is given in [Lipton82], [Dodgson05] and [Zilly13].

2.1.1.3 Multi-Projector Lightfield Displays

The stereoscopic and auto-stereoscopic displays presented above are based on an underlying 2D panel, hence the total resolution is limited to the respective panel resolution, i.e. 2k or 4k. However, this resolution needs to be shared among all views. Another approach is followed by [Iwasawa11] and [Balogh07]. The former uses an array of 57 Full HD projectors along with a condenser lens and a special diffusion screen to allow a high definition auto-stereoscopic viewing experience on a 200"

⁶ INFITEC GmbH, 3D-Visualisierungstechnik. Website: <http://www.infitec.net/>

screen. The principle of [Balogh07] is similar. Here, up to 80 HD projectors are used to create a 140" wide auto-stereoscopic screen. Due to the high number of views, the amount of reproducible parallax is high which requires multi-camera arrays to capture content for these displays.

Many other 3D display techniques have been proposed in the past. Moreover, new concepts are constantly developed and presented to the public. A recent example is given in [Nagano13]. A good overview of different 3D display technologies is given in [Holliman11].

2.1.2 Basic Geometric Concepts

As described in the previous sections, the underlying concept of 3D displays is to offer two or more different views and perspectives. Different views are thereby offered to the left and right eye of the observer. In this case, the perception of binocular depth cues (see sub-section 2.2.1) results from the spatial distances between corresponding points in both planar views, i.e., from the so-called screen disparity or parallax \mathcal{P} , which in turn induces the retinal disparities in the viewer's eyes (see sub-section 2.2.1). In Figure 2.3 three different cases for the parallax \mathcal{P} are illustrated. The parallax can be either positive, negative, or zero.

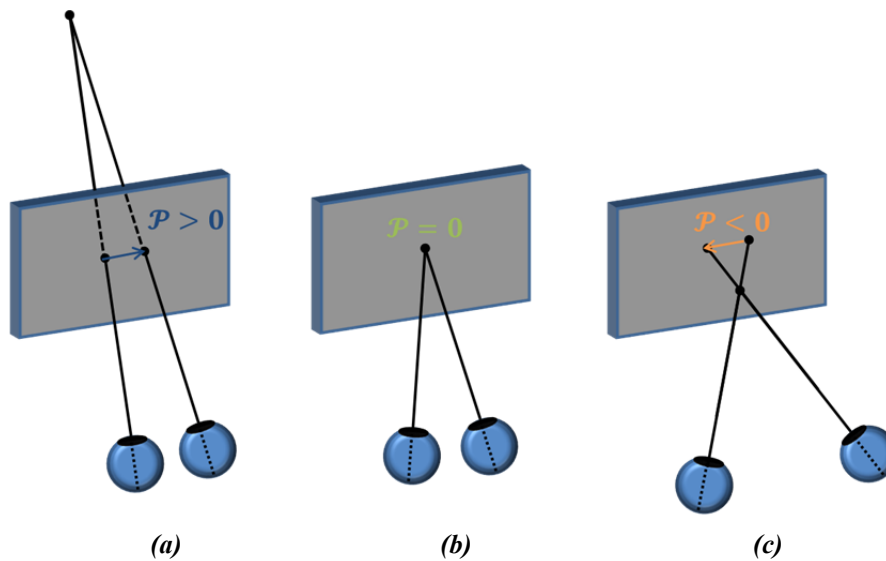


Figure 2.3. Different types of parallax: (a) positive parallax visualized in blue; (b) zero parallax visualized in green; (c) negative parallax visualized in orange.

In Figure 2.3 (a), an example for a positive parallax is shown. As a result, the eyes converge at a distance behind the screen in the so-called screen space. If the parallax would exceed the inter-ocular distance t_{eye} , the eyes would diverge. In Figure 2.3 (b), the parallax is zero, i.e. corresponding points are seen at the same position on the screen which means that an object is perceived at the screen distance. In Figure 2.3 (c), the parallax is negative, the eyes converge in front of the screen in the so-called viewer space.

These schemata can now be used to derive a mathematic expression for the reconstructed depth Z_v in relation to the parallax \mathcal{P} , the viewing distance Z_D and inter-ocular eye-distance t_{eye} .

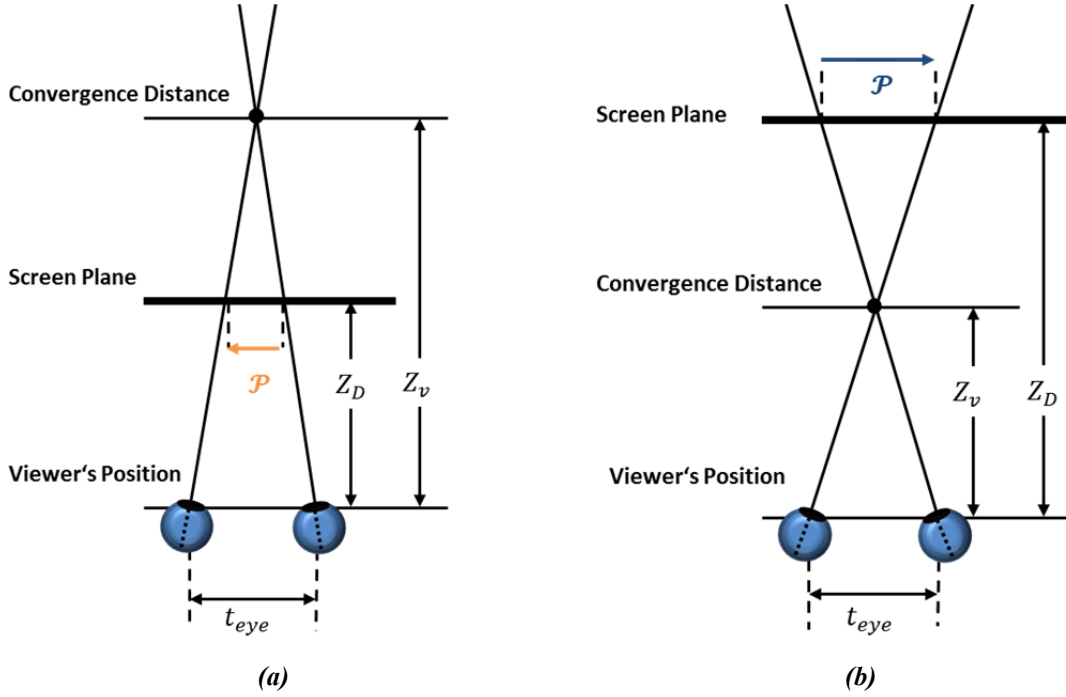


Figure 2.4. Illustration of geometrical relation described in eqns. (2.1) and (2.2). The convergence distance of the eyes and the resulting perceived distance Z_v is a function of the screen parallax \mathcal{P} , the distance to the screen Z_D and the inter-ocular distance t_{eye} .

Given the geometric setup shown in Figure 2.4 and by referring to the intercept theorem, the relation between the mentioned geometric entities is given by the following equation:

$$\frac{\mathcal{P}}{t_{eye}} = \frac{Z_v - Z_D}{Z_v}. \quad (2.1)$$

The same relation holds true for both cases (a) and (b) of Figure 2.4 as well as for the case where the screen plane and the convergence plane coincide, i.e. if $Z_D = Z_v$ and $\mathcal{P} = 0$. The equation (2.1) can be reformulated to express the reconstructed depth Z_v of an object which appears on the screen at a viewing distance Z_D with given parallax \mathcal{P} :

$$Z_v = \frac{Z_D \cdot t_{eye}}{t_{eye} - \mathcal{P}}. \quad (2.2)$$

The eqn. (2.2) can be used to describe the geometrical implications of the different types of parallax illustrated in Figure 2.3:

- If $\mathcal{P} < 0$, it follows that $Z_v < Z_D$, i.e. the object is seen in front of the screen;
- If $\mathcal{P} = 0$, it follows that $Z_v = Z_D$, i.e. the object is seen at the screen;
- If $t_{eye} > \mathcal{P} > 0$, it follows: $Z_v > Z_D$, i.e. the object is behind the screen at a finite distance;
- If $\mathcal{P} = t_{eye}$, the denominator of eqn. (2.2) vanishes, Z_v equals to infinity;
- If $\mathcal{P} > t_{eye}$, it follows that $Z_v < 0$, which is an unnatural viewing condition, the eyes of the viewer need to diverge.

2.1.3 Horizontal Image Translation and Parallax Range

As described in eqn. (2.2) the distance Z_v at which an object is visualized on a 3D screen depends on the parallax \mathcal{P} , beside other parameters such as the inter-ocular distance t_{eye} and the viewing distance Z_D .

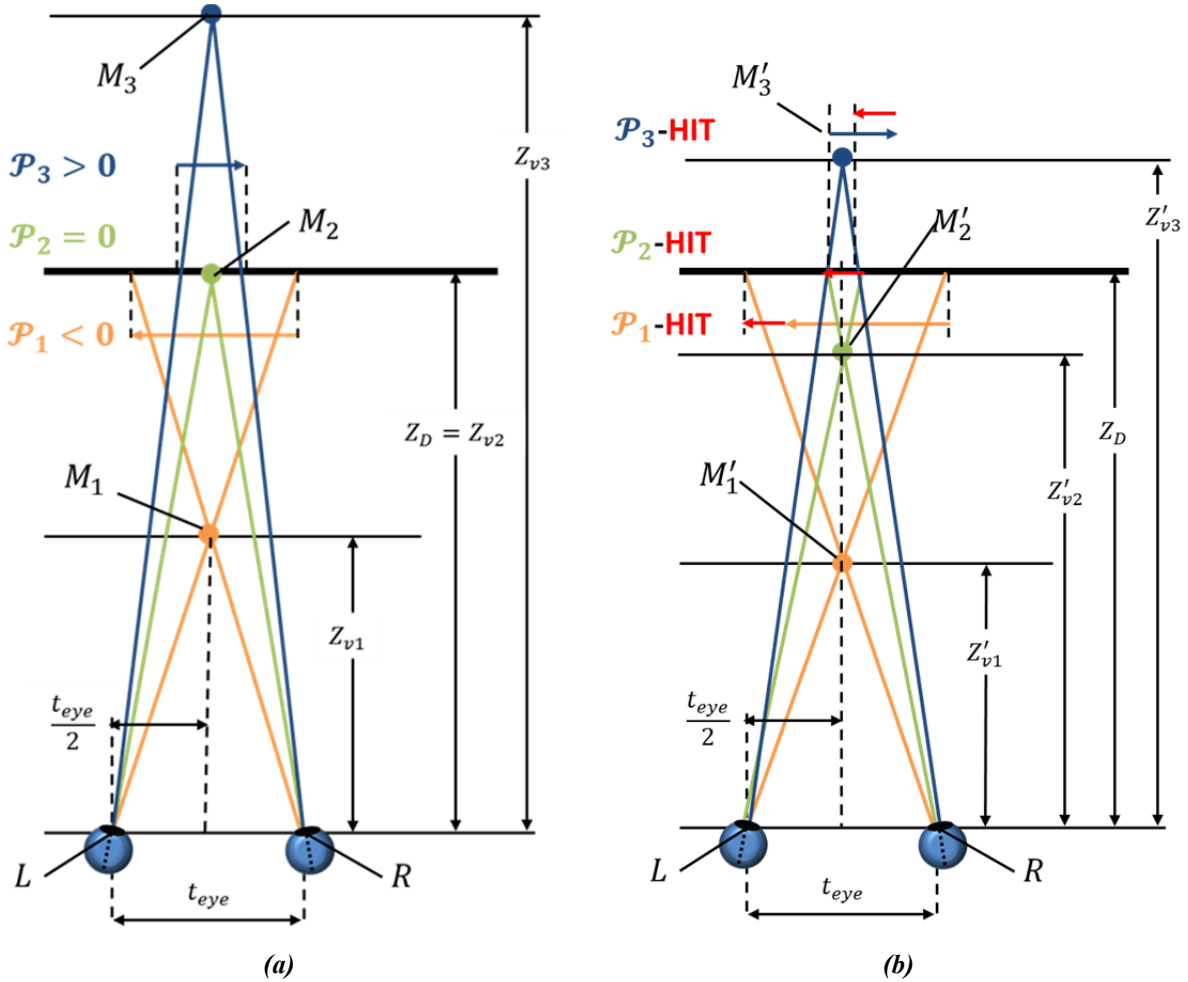


Figure 2.5. (a) The three points M_1 , M_2 and M_3 are reproduced according to their parallax values \mathcal{P}_1 , \mathcal{P}_2 and \mathcal{P}_3 . (b) By adding an offset, or horizontal image translation (HIT), it is possible to shift the parallax in positive or negative direction. The HIT leads to a change of the perceived distances, e.g. Z_{v1} shifts to Z'_{v1} etc.

The relation is illustrated in Figure 2.5. Three points M_1 , M_2 and M_3 are reproduced at the distances Z_{v1} , Z_{v2} and Z_{v3} according to their parallax values \mathcal{P}_1 , \mathcal{P}_2 and \mathcal{P}_3 . It is possible to apply a horizontal image translation (HIT) which acts as an offset for the parallax values as shown in Figure 2.5(b). Although the parallax offset for the three points can be changed, it is not possible to modify the parallax range, e.g. $\mathcal{P}_3 - \mathcal{P}_1$ remains constant regardless of the choice of the HIT.

2.2 The Human Visual System and Depth Perception

In this section, an overview of the principles of the human visual system (HVS) and its mechanism to perceive depth is given⁷. Our brain is able to interpret several types of depth cues which can be separated into two categories: monocular and binocular depth cues [Lipton82]. Monocular cues require only one eye, while binocular cues involve both eyes. Typical scenes stimulate several depth cues at once in the HVS. The interpretation of the cues is based on learning and experience [Lipton97]. The relevance of the different cues for the human depth perception depends according to Cutting on the relative distances between the observer and the objects in the scene. Cutting proposes to differentiate between personal space, action space, and vista space [Cutting97]. As the binocular cues are of particular importance for stereoscopic viewing, the following sub-section starts with a description of these depth cues, followed by a further sub-section on the monocular depth cues that provide the human brain with additional depth information and are often the reason for the occurrence of unwanted perception conflicts during 3D stereo reproduction.

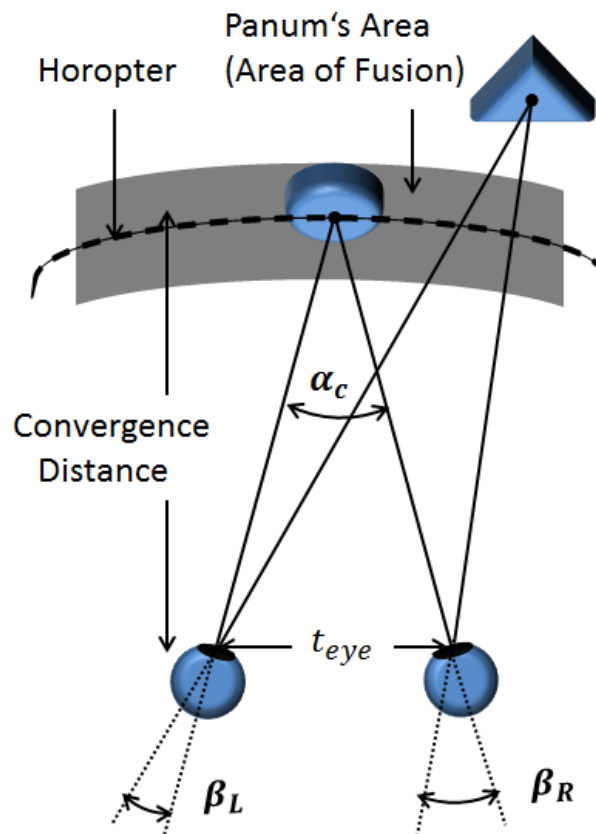


Figure 2.6. Principal of convergence and retinal disparity. By forming the angle α_c the eyes are converged on the disk, which is imaged in the centers of both retinas. The triangle in contrast is projected to different positions described by the corresponding angles β_L and β_R . Retinal disparity is a function of the difference of the angles $\Delta\beta = \beta_R - \beta_L$. (Drawing based on figures adapted from [Lipton82], [Jsselsteijn02] and [Zilly2011b]).

⁷ Parts of the content in this section have been previously published in [Zilly11b].

2.2.1 Binocular Depth Cues

Binocular depth cues take advantages of the spatial separation of the human eyes, i.e. the mainly horizontal stereo baseline called inter-ocular eye distance t_{eye} . It has a value of approximately 64 mm for an adult [Lipton82] and is visualized in Figure 2.6. The two viewing positions create two unique perspectives of the observed scene.

The two binocular depth cues are *convergence* and *retinal disparity*. In the following, the geometric principle of these depth cues is described.

2.2.1.1 Convergence

As illustrated in Figure 2.6, the two eyes rotate in order to adapt the convergence distance to the object of interest (i.e. the disk). The angle between the two intersecting optical axes is the convergence angle α_c . By triangulation, the HVS is able to calculate the distance between the viewer and the object at convergence distance [Lipton82].

2.2.1.2 Retinal Disparity

As illustrated in Figure 2.6, the two objects, i.e. the disk and the triangle are projected onto different positions in the retina. The differences can be described by the angles β_L in the left eye and β_R in the right eye under which the optical rays intersect. The retinal disparity is a function of the difference of the angles, i.e. $\Delta\beta = \beta_R - \beta_L$ which is zero for all objects on the so-called Horopter (see Figure 2.6) [Lipton82]. The small zone around the Horopter is called Panum's Area, and objects within this zone can be directly fused by the HVS [Lipton82, IJsselstein02, Lambooi09].

2.2.2 Monocular Depth Cues

Beside the binocular depth cues, which use both eyes there are a different monocular depth cues which allow the HVS to extract depth information using a single eye only. For far distances, the monocular depth cues can even be the dominant source for depth information [Cutting97]. A non-exhaustive list of monoscopic depth cues is presented in the following:

- Accommodation and blur: The eyes accommodate their focus to a distance at which objects are seen sharply – in contrast to blurred objects, which allows judging the distance of the objects.
- Relative size: When the typical size of an object such as a car or a pedestrian in a scene is known, one can estimate their distance from the viewer.
- Interposition: One object which occludes a second object is obviously nearer to the viewer.
- Motion Parallax: When the observer moves without turning the head, nearer objects cross faster the visual field than farther objects. An effect which can be well observed when watching through the windows of a moving train.

More details on monocular depth cues can be found in [Lipton82] and [Zilly2011b].

2.2.3 3D Perception Conflicts

Ideally, the different depth cues deliver consistent information to the human visual system. Conflicting depth cues in contrast, can cause 3D perception conflicts as well as unnatural viewing conditions, e.g. when the eyes need to diverge. In the following, a selection of important 3D perception conflicts is presented:

- Binocular Rivalry and Vertical Disparities,
- Accommodation-Convergence Conflict,
- Stereo Framing,
- Eye Divergence.

2.2.3.1 Binocular Rivalry

Binocular or retinal rivalry denotes a phenomenon of the human visual system which occurs when inconsistent information processed by monocular cues in the left and right eye shall be fused by the brain into a single image [Lipton82]. A significant luminance difference is one of many possible sources for the phenomenon [Beldie91]. Similarly, inconsistent contrast levels can lead to binocular rivalry [Dumbreck98]. Moreover, due to imperfections of the 3D reproduction chain, it might occur that the channel separation between the left and the right images is impaired [Lipton01] which yields to an effect called cross-talk or ghosting [Mendiburu2008]. Technical reasons are for instance cross-talk between pixels of a polarizing 3D-TV screen. As a result, the spectator sees in the left (right) eye information not only from the left (right) stereo channel, but also from the right (left) stereo channel. These so-called ghost images impair the 3D sensation. The effect is usually stronger if high contrasts exist in the stereo images. To avoid this undesired effect, the stereographer tries to avoid high contrasts at least for objects which are not in the convergence plane where no ghost images can occur as objects belonging to the left and right view are reproduced at the same position on the screen [Mendiburu08]. The phenomenon of binocular rivalry was first investigated in detail by Wheatstone [Wheatstone38] and was since then topic of intensive research. An extensive description of underlying mechanism along with an overview of state-of-the art research results is given in [Alais05]. Binocular rivalry can lead to a temporal loss of 3D sensation [Lipton82]. Similarly to binocular rivalry, vertical disparities resulting from misalignment of the cameras or lens distortion can impair the 3D sensation and lead to eye-strain [Woods93, Zilly11b].

2.2.3.2 Stereo-Framing / Window Violation

When an object with negative parallax \mathcal{P}_{object} is cut-off by the border of the 3D screen, a so-called stereo *window violation* occurs [Dashwood10]. The border of the 3D screen has the same parallax as the screen itself, i.e. $\mathcal{P}_{border} = 0$ according to eqn. (2.2). This induces a conflict between the binocular depth cue *retinal disparity* and the monocular depth cue *interposition*. The effect is visualized in Figure 2.7.

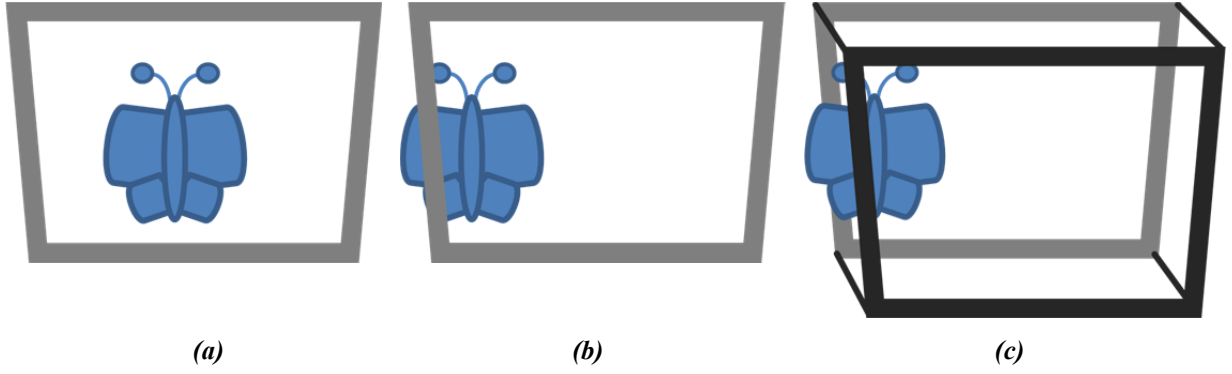


Figure 2.7. (a) The object does not interfere with the frame border, hence no conflict occurs. (b) The object with parallax $\mathcal{P}_{object} < 0$ is occluded by the frame border with parallax $\mathcal{P}_{border} = 0$ which causes a conflict between the depth cues retinal disparity and interposition. (c) A floating window with parallax $\mathcal{P}_{floating} < \mathcal{P}_{object}$ is inserted which solves the conflict.

The conflict can be solved by changing the convergence plane using a horizontal image translation (HIT) or by inserting a floating window as shown in Figure 2.7(c). The floating window has negative parallax and is inserted into the footage by adding black bars at different positions in the left and right view [Mendiburu2012]. The aim is to create the illusion, that the floating window with parallax $\mathcal{P}_{floating} < \mathcal{P}_{object}$, which is nearer to the viewer, occludes the foreground object (i.e. the butterfly in Figure 2.7(c)). Consequently, the conflict between the depth cues *retinal disparity* and *interposition* is solved or at least reduced.

2.2.3.3 The Accommodation-Convergence Conflict

In a natural environment, the eyes focus, or accommodate to the distance onto which the eyes converge on, i.e. the convergence distance coincides with the accommodation distance as shown in Figure 2.8 (a).

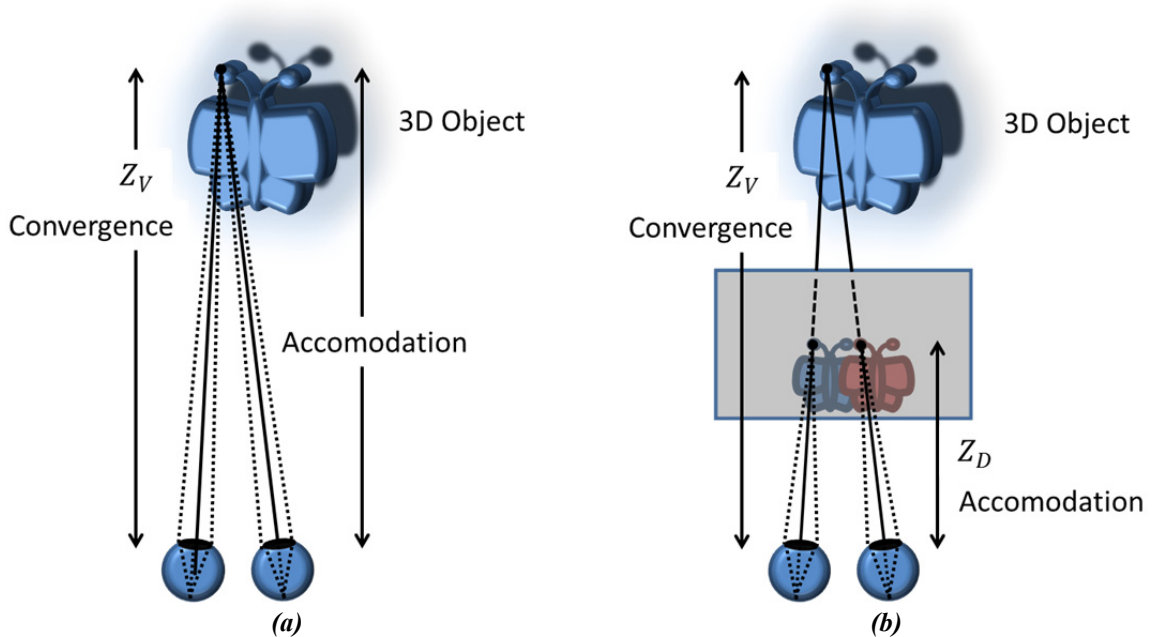


Figure 2.8. Conflict between accommodation and convergence when looking at a stereoscopic 3D display.

The situation differs in the case of a stereoscopic 3D reproduction where the eyes accommodate to the screen surface at distance Z_D while the convergence distance Z_V depends on the screen parallax \mathcal{P} as depicted in Figure 2.8 (b). Consequently, the depth cues related to accommodation and convergence deliver inconsistent information to the human visual system. This results in the so-called accommodation-convergence conflict which can lead to a loss of the 3D sensation [Lipton82, Lambooi09].

Lipton describes in [Lipton82] that a complete loss of 3D sensation is given when the screen parallax \mathcal{P} reaches 3% of the viewing distance Z_D according to the illustration from Figure 2.4. The limit for a comfortable viewing sensation which can be expressed as the ratio between screen parallax L_{AC} is object of research. For the retinal disparities, a value of 70 arc minutes is described in the literature [Lambooi09], [Pastoor95], [Wopking95] which translates to a ratio between screen parallax and viewing distance of approximately 2%, i.e. $L_{AC} \approx 2\%$.

2.2.3.4 Eye Divergence

In a natural environment, the eyes converge to an object at converge distance. As discussed earlier in section 2.1.2, this concept, can be applied to a 3D display where the screen parallax, in combination with the viewing distance, defines the convergence plane as visualized in Figure 2.9 (a). In contrast, unnatural viewing conditions are given according to eqn. (2.2) if the screen parallax \mathcal{P} exceeds the inter-ocular eye distance t_{eye} as shown Figure 2.9 (b).

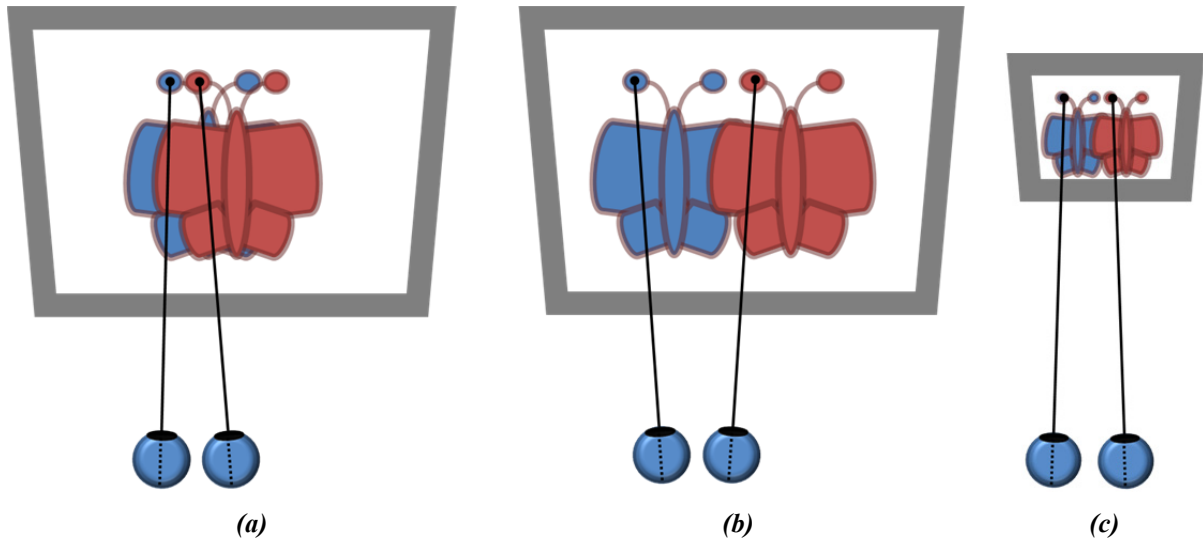


Figure 2.9. (a) The eyes converge to a finite convergence distance. (b) The eyes diverge as the screen parallax exceeds the inter-ocular eye distance. (c) Although the same content as is (b) shown, the eyes converge due to the smaller screen which leads to a smaller screen parallax.

However, the same content reproduced on a smaller screen as illustrated in Figure 2.9 (c), might not lead to eye divergence.

2.3 Geometrical Implications for Stereoscopic 3D Production

2.3.1 Comfortable Viewing Range and Depth Budget

According to Sun and Holliman [Sun2009], the perceived depth should be limited to a *comfortable viewing range* (CVR). The meaning is similar to the term *depth budget* [Mendiburu2008, Mendiburu2012]. The CVR takes into account viewing conditions such as display size and viewing distance. The depth budget is thereby a result of the considerations which aim at avoiding 3D perception conflicts as described in section 2.2.3. The excerpt of the considerations can be summarized as follows:

Rule 1: $\mathcal{P} > \mathcal{P}_{floating}$ for objects interfering with the frame border to avoid window violation. (2.3)

Rule 2: $-L_{AC} < \mathcal{P}/Z_D < L_{AC}$, with $L_{AC} \approx 2\%$ to avoid accommodation-convergence conflict. (2.4)

Rule 3: $\mathcal{P} \leq t_{eye}$, to avoid eye divergence. (2.5)

The aim is now to calculate an upper limit \mathcal{P}_{max} and a lower limit \mathcal{P}_{min} for the parallax which take into account the rules from eqns. (2.3)-(2.5). It is usual to express the parallax limits as a percentage of the screen width $\tilde{\mathcal{P}}$. Against this background, the screen parallax can be normalized with respect to the screen width W_{screen} :

$$\tilde{\mathcal{P}} = \frac{\mathcal{P}}{W_{screen}}, \tilde{\mathcal{P}}_{max} = \frac{\mathcal{P}_{max}}{W_{screen}}, \tilde{\mathcal{P}}_{min} = \frac{\mathcal{P}_{min}}{W_{screen}} \quad (2.6)$$

The above considerations can be used to define the depth budget depending on the viewing conditions in a quantitative way. An example is given in Table 2.1 which originates from [Knorr12]. According to Knorr et al. the limits for the negative parallax and positive parallax depend on the screen size. The values from Table 2.1 are a very compact representation format for defining the position and the width of the comfortable viewing range.

Table 2.1. Typical Depth Budgets of different screen types according to [Knorr12].

Display Type	Near Limit: Negative Parallax $\tilde{\mathcal{P}}_{min}$	Far Limit: Positive Parallax $\tilde{\mathcal{P}}_{max}$
IMAX	1.5-2.5%	<0.25%
Cinema	2%	<1%
TV	1-1.5%	1.5-2%

As a rule of thumb, Mendiburu et al. propose values of 2% of negative parallax and 2% of positive parallax as useful limits ([Mendiburu2012], p. 170). The ideal depth budget might be smaller, especially for large screens, while on the other hand it might be allowed to exceed the depth budget for a short amount of time [Zilly2011b].

2.3.2 Geometrical Concepts of 3D Acquisition

The generation of stereoscopic views requires the capture of a real 3D scene with two aligned stereo cameras as introduced in section 1.2.1. Thereby, a side-by-side configuration or a beam-splitter as illustrated earlier in Figure 1.1 can be used. In this section, another important aspect of the 3D acquisition geometry shall be described⁸. The setup of the stereo cameras can be either convergent or parallel as shown in Figure 2.10.

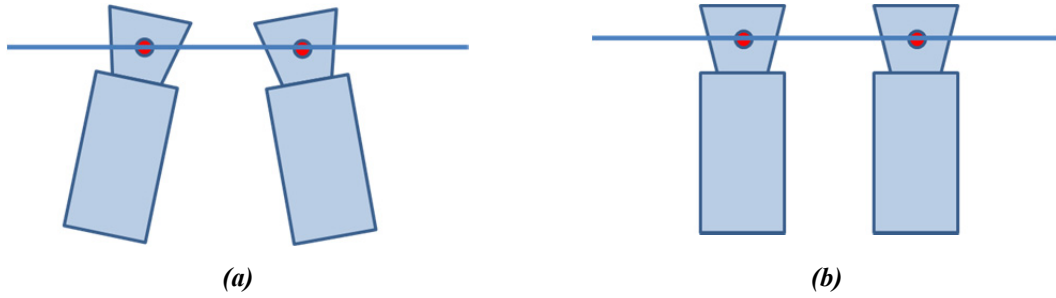


Figure 2.10. Basic stereoscopic camera configurations: (a) convergent or toed-in approach; (b) parallel setup.
When using the convergent approach, the convergence distance has a finite value but vertical disparities are introduced which can be eliminated by stereo image rectification. When using the parallel approach, no vertical disparities occur, but the convergence distance is at infinity, hence the images have to be shifted electronically using a horizontal image translation (HIT) prior to the reproduction on a 3D display.

Both camera configurations have advantages and disadvantages concerning the 3D production workflow. When using the convergent approach as shown in Figure 2.10 (a), the convergence distance has a finite value but vertical disparities are introduced [Woods93] which can be eliminated by applying a stereo image rectification (see section 2.4.2.2 for details on rectification). When using the parallel approach as shown in Figure 2.10 (b), no vertical disparities occur, but the convergence distance is at infinity, hence the images have to be shifted electronically using a horizontal image translation (HIT) prior to the reproduction on a 3D display.

For the parallel camera setup, the horizontal disparity d which denotes the displacement (in pixels) of corresponding pixels in the left and right view can be calculated using the following formula:

$$d = B \cdot \frac{f}{Z} \quad (2.7)$$

which is derived in section 2.4.2.2. Here, f , Z , and B denote the focal length of the two stereo cameras, the distance of the world point from the cameras and the stereo baseline, respectively.

If the distance of the nearest and the farthest objects in a 3D scene is known, it is possible to calculate the disparity range Δd as follows:

$$\Delta d = d_{max} - d_{min} = B \cdot f \cdot \left(\frac{1}{Z_{near}} - \frac{1}{Z_{far}} \right). \quad (2.8)$$

⁸ Parts of the content in this section have been previously published in [Zilly11b].

where d_{max} and d_{min} denote the disparities of the farthest and nearest objects respectively. The disparity d which can be measured in pixels can also be expressed in a resolution-agnostic way, e.g. by normalizing it with respect to the sensor width W_{sensor} , i.e. 1920 pixels in the case of full HD sensors:

$$\tilde{d} = d/W_{sensor}. \quad (2.9)$$

Similarly, the disparity range can be expressed in a resolution-agnostic way:

$$\Delta\tilde{d} = \tilde{d}_{max} - \tilde{d}_{min} = d_{max}/W_{sensor} - d_{min}/W_{sensor} = \mathcal{B} \cdot f/W_{sensor} \cdot \left(\frac{1}{z_{near}} - \frac{1}{z_{far}} \right). \quad (2.10)$$

Assuming that the entire image which is recorded by the cameras will be shown on the screen, one can deduce that the relative parallax ranges from Table 2.1 equal to the relative disparity range:

$$\Delta\tilde{d} = \Delta\tilde{\mathcal{P}} = \tilde{\mathcal{P}}_{max} - \tilde{\mathcal{P}}_{min}. \quad (2.11)$$

2.3.3 Adjustment of Inter-Axial Camera Distance

It is the task of the stereographer to ensure that the inter-axial distance is chosen such that the depth volume is within a comfortable range.

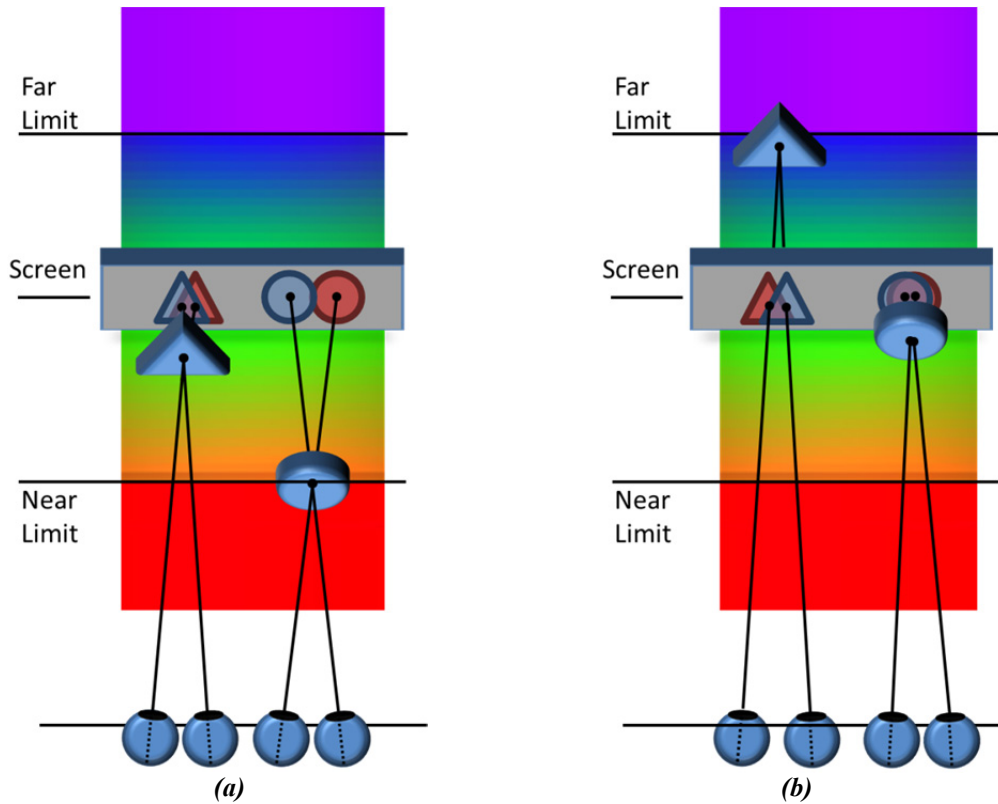


Figure 2.11. The comfortable viewing range can be visualized as green zone around the screen plane. Behind the far limit the positive parallax is exceeded (colored in violet). In front of the near limit, the minimum parallax is exceeded (colored in red).

When the allowed depth budget and the scene to be shot is known, the following formula can be used to calculate a suitable stereo baseline \mathcal{B} :

$$B = \frac{\tilde{d}_{max} - \tilde{d}_{min}}{f \cdot \left(\frac{1}{Z_{near}} - \frac{1}{Z_{far}} \right)} = \frac{\tilde{\mathcal{P}}_{max} - \tilde{\mathcal{P}}_{min}}{f \cdot \left(\frac{1}{Z_{near}} - \frac{1}{Z_{far}} \right)}. \quad (2.12)$$

The aim is thereby to ensure that all objects in the scene can be reproduced within the comfortable viewing range, i.e. within the far and near limit as illustrated in Figure 2.11.

A detailed comparison between traditional and a proposed assisted adjustment of the inter-axial distance is performed in section 5.5.3.

2.3.4 Adjustment of the Convergence Distance and Horizontal Image Translation

It is the task of the convergence puller to ensure that a proper convergence distance is chosen during the 3D production process. As illustrated in Figure 2.11, the different convergence planes might be suitable as long as the whole scene can be reproduced within the comfortable viewing range. As shown in section 2.1.3, a horizontal image translation (HIT) can be applied to add an offset to the minimal and maximal parallax values.

A detailed comparison between traditional and a proposed assisted adjustment of the convergence distance is performed in section 5.6.2.

2.3.5 Mechanical Alignment and Setup of the 3D Rig

The task of the stereographer or camera assistance is to ensure that the mechanical alignment of the stereo cameras is ensured. A detailed comparison between traditional and a proposed assisted rig calibration workflow is performed in section 5.6.1.

Beside the mechanical alignment, the stereographer has to ensure a temporal synchronization of the stereo cameras. If the left and the right cameras would capture the images at different instances in time, moving objects would induce additional horizontal and vertical disparities depending of the nature of the motion which impairs the 3D perception of the stereoscopic content [Lipton82]. Another inconsistency of the left and right images occurs if the exposure times are not identical. In this case, objects in the left and the right cameras are affected by different motion blur. Last but not least, the depth of field seen in the left and right camera need to match in order to avoid varying sharpness for objects seen in the left and right camera. Thus, the focus distance and the aperture need to match. The latter does not only have an impact to the brightness, but also to the sharpness of the depth of field, which should be symmetric [Routier12].

2.4 Projective Geometry

In this section, basic concepts of the projective geometry are introduced⁹. A good overview of this field of research is given in [Faugeras93b] and [Hartley04]. Images captured by a camera are

⁹ Parts of the content in this section have been previously published in [Zilly10a] and [Zilly12c].

projections of the 3D world onto a 2D surface. A common camera model used in computer vision is the pinhole camera. The pinhole camera can be thought as an empty cuboid object with a small hole permitting the light to enter and to be projected on the rear surface of the cuboid as illustrated in Figure 2.12. As can be seen, all light rays have to pass the pinhole and get projected onto the 2D surface forming an upside down image of the 3D object.

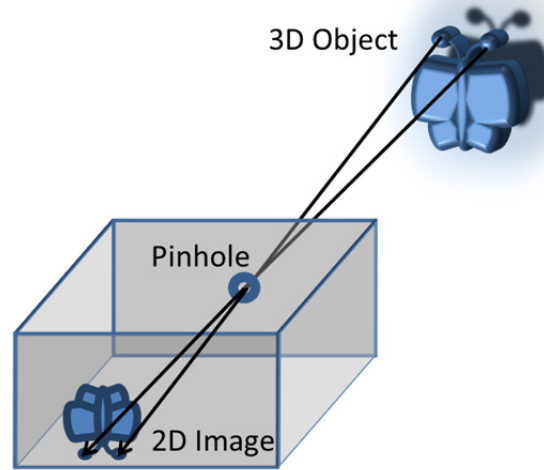


Figure 2.12. *Concept of the pinhole camera. The 2D image is affected by projective distortions.*

Within the idealized concept of the pinhole camera, the pinhole is infinitely small, i.e. all light rays have to pass through the exact same point. In reality however, the pinhole has to have a finite size. In fact, the smaller the pinhole becomes, the fewer light rays can enter the box. The wider the pinhole is, the more light enters the camera, but the blurrier the images becomes. However, the image becomes also blurry due to diffraction effects, if the pinhole is too small [Bergmann93].

To overcome the limitations of the antagonistic effect of sharpness and luminance, light collecting lenses are used in modern cameras, sometimes introducing other image distortions effects such as radial lens distortion, coma, chromatic aberration, and other effects such as depth of field [Bergmann93]. Nevertheless, in order to simplify further calculations, it can be assumed that all these effects can either be neglected or corrected digitally. Hence, a camera can be thought as an ideal pinhole camera. As mentioned above, during the acquisition of an image, 3D world coordinates are projected onto 2D coordinates which usually consist of the pixel positions of the camera sensor.

2.4.1 Basic Camera Geometry

As shown in Figure 2.12, the image of the object is mirror-inverted. However, by a simple rotation of 180° around the image center, one can recover the more intuitive image orientation. In fact, this simple process is performed automatically in digital cameras.

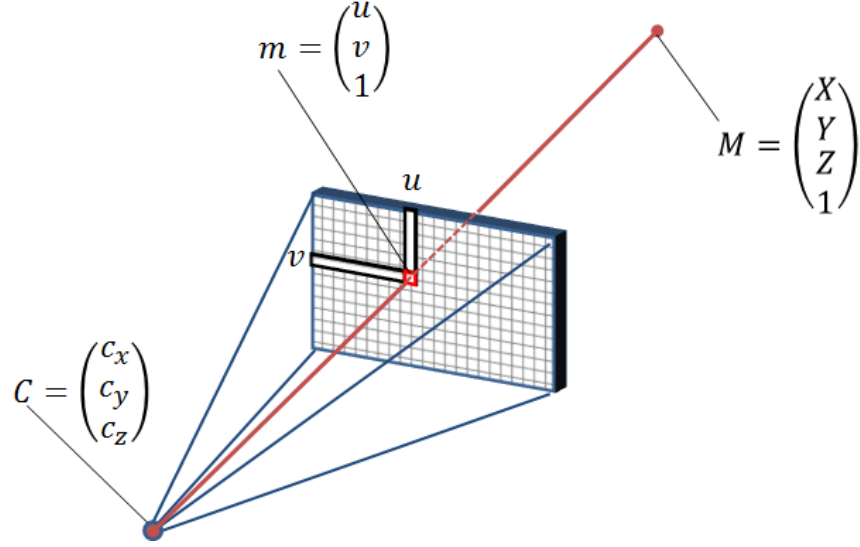


Figure 2.13. Projection of a world point M onto an image sensor.

In the following, a mathematical formulation of this process will be given. In Figure 2.13, the concept of projecting a 3D world point onto two-dimensional image coordinates is illustrated. In contrast to Figure 2.12, the sensor is drawn in between the world point and the camera sensor while the above rotation of 180° of the sensor has already been performed. Homogeneous coordinates were assigned to the 3D world point M which facilitates to express not only rotations but also translations by specific matrices representing the geometry:

$$M = \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}. \quad (2.13)$$

As shown in Figure 2.13, a red line is drawn between M and the camera center C which can be thought as the pinhole from Figure 2.12. The line represents a light ray which hits the image sensor made of rectangular pixels at the pixel position m with coordinates u , v , and 1. The pixel position m at which the light ray hits the sensor depends obviously of the position of the object assigned to the 3D world point M but also on the position of the camera center C , the orientation of the camera, and intrinsic camera parameters such as the focal length. In the following, the geometric parameters will be discussed and a mathematical formulation of the projection will be given.

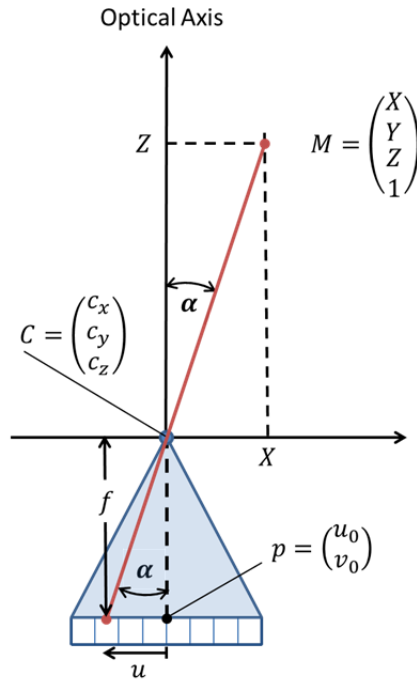


Figure 2.14. Geometry of a light ray projected onto a camera.

Figure 2.14 illustrates the geometry of the projection process. The world coordinate system has been rotated such that the optical axis which is perpendicular to the image sensor and passes through the pinhole and camera center \mathbf{C} coincides with the z -axis while the x -axis is parallel to the image sensor lines. The projection of the optical axis onto the image sensor is the principal point \mathbf{p} with coordinates u_0 and v_0 . The distance between image sensor and camera center is denoted as focal length f . In the first step, z -coordinates are ignored in order to simplify the calculation. From intercept theorem it is known that the following ratios are equal:

$$\frac{X - c_x}{Z - c_z} = \frac{u - u_0}{f}. \quad (2.14)$$

Eqn. (2.14) is rewritten in order to calculate the horizontal component u of the pixel position of the projected 3D world point:

$$u = f \frac{X - c_x}{Z - c_z} + u_0. \quad (2.15)$$

An analog calculation can be performed when introducing the y -components:

$$v = f \frac{Y - c_y}{Z - c_z} + v_0. \quad (2.16)$$

As mentioned above, the usage of homogeneous coordinates allows expressing the projection calculus using matrix multiplications. The results from eqns. (2.15) and (2.16) can be converted into the

following equation yielding to the same results where s is a scaling factor for the homogeneous coordinates:

$$\underbrace{\begin{pmatrix} s \cdot u \\ s \cdot v \\ s \end{pmatrix}}_{\mathbf{m}} = \underbrace{\begin{pmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{K}_s} \cdot \underbrace{\begin{pmatrix} 1 & 0 & 0 & -c_x \\ 0 & 1 & 0 & -c_y \\ 0 & 0 & 1 & -c_z \end{pmatrix}}_{\mathbf{I} | -\mathbf{c}} \cdot \underbrace{\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}}_{\mathbf{M}}. \quad (2.17)$$

Please note that in the above equation, it is assumed that the origin of the pixel coordinates is in the image center. As can be seen, the first and the second components of \mathbf{m} have to be divided by the third component s , in order to get the final result for u and v .

Until now, it was assumed that the cameras' optical axis coincides with the z -axis of the world coordinate system and that the pixels are square, while each image scanline coincides with the x -axis of the world coordinate system.

In the general case, the orientation of the camera with respect to the world coordinate system is arbitrary and can be described by a corresponding rotation matrix $\mathbf{R}(\alpha_x, \alpha_y, \alpha_z)$ defined as follows:

$$\mathbf{R}(\alpha_x, \alpha_y, \alpha_z) = \begin{pmatrix} \cos(\alpha_y)\cos(\alpha_z) & \sin(\alpha_x)\sin(\alpha_y)\cos(\alpha_z) - \cos(\alpha_x)\sin(\alpha_z) & \sin(\alpha_x)\sin(\alpha_z) + \cos(\alpha_x)\sin(\alpha_y)\cos(\alpha_z) \\ \cos(\alpha_y)\sin(\alpha_z) & \sin(\alpha_x)\sin(\alpha_y)\sin(\alpha_z) + \cos(\alpha_x)\cos(\alpha_z) & \cos(\alpha_x)\sin(\alpha_y)\sin(\alpha_z) - \sin(\alpha_x)\cos(\alpha_z) \\ -\sin(\alpha_y) & \sin(\alpha_x)\cos(\alpha_y) & \cos(\alpha_x)\cos(\alpha_y) \end{pmatrix}. \quad (2.18)$$

The intrinsic camera parameters skew, and the pixel aspect ratio need to be added to the position of the principal point $\mathbf{p}(u_0, v_0)$ and the focal length f . The skew vanishes if the rows and columns of the image sensor are perpendicular, which is often assumed to be the case. Nevertheless, a slanted pixel raster can be described by this parameter. The pixel aspect ratio can also be interpreted as the ratio of the horizontal focal length f_x to the vertical focal length f_y . The intrinsic matrix \mathbf{K} is presentend in eqn. (2.19):

$$\mathbf{K} = \begin{pmatrix} f_x & skew & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (2.19)$$

2.4.1.1 Projection Matrix

Enhancing eqn. (2.17) by the rotation matrix \mathbf{R} and the updated intrinsic matrix \mathbf{K} from (2.19) yields to the following equation:

$$\underbrace{\begin{pmatrix} s \cdot u \\ s \cdot v \\ s \end{pmatrix}}_{\mathbf{m}} = \underbrace{\begin{pmatrix} f_x & skew & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{K}} \cdot \underbrace{\mathbf{R}(\alpha_x, \alpha_y, \alpha_z)}_{\mathbf{R}} \cdot \underbrace{\begin{pmatrix} 1 & 0 & 0 & -c_x \\ 0 & 1 & 0 & -c_y \\ 0 & 0 & 1 & -c_z \end{pmatrix}}_{\mathbf{I} | -\mathbf{c}} \cdot \underbrace{\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}}_{\mathbf{M}}. \quad (2.20)$$

Eqn. (2.20) describes the projection of a 3D world point onto 2D image coordinates in the general case. The components which parameterize the projection can be regrouped to the following 4x3 matrix which is called the projection matrix \mathbf{P} :

$$\mathbf{P} = \mathbf{K}\mathbf{R} [\mathbf{I} | -\mathbf{C}]. \quad (2.21)$$

Using eqn. (2.21), one can rewrite eqn. (2.20) in a more compact way:

$$\mathbf{m} = \mathbf{P} \cdot \mathbf{M}. \quad (2.22)$$

The above equation describes the projection of a 3D world point onto 2D image coordinates.

2.4.1.2 Projective Transformations and Homographies

The mapping of a 3D world point onto 2D image coordinates can be expressed by the projection matrix as defined in eqn. (2.21). However, one can imagine that a second camera with the same camera center \mathbf{C} , or pinhole, but different orientation or intrinsic parameters maps the 3D world point onto different pixel coordinates \mathbf{m}' . In this case, there is an invertible relationship between the corresponding pixel coordinates \mathbf{m} and \mathbf{m}' called projective transformation or homography \mathbf{H} [Hartley04] with:

$$\mathbf{m}' = \mathbf{H} \cdot \mathbf{m}, \quad (2.23)$$

where \mathbf{H} is a non-singular 3x3 matrix:

$$\mathbf{H} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{pmatrix}. \quad (2.24)$$

In the context of this thesis, homographies will be used to normalize cameras parameters and to perform stereo or multi-camera rectifications.

2.4.2 Two-Camera Geometry

In the previous sub-section, important elements of the projective geometry using a single camera were introduced. This sub-section focusses on the projective geometry with two cameras which plays an important role for stereoscopic 3D productions. An introduction to the epipolar geometry and the fundamental matrix \mathbf{F} will be given. Subsequently, the concept of stereo-rectification will be explained along with an overview of related work.

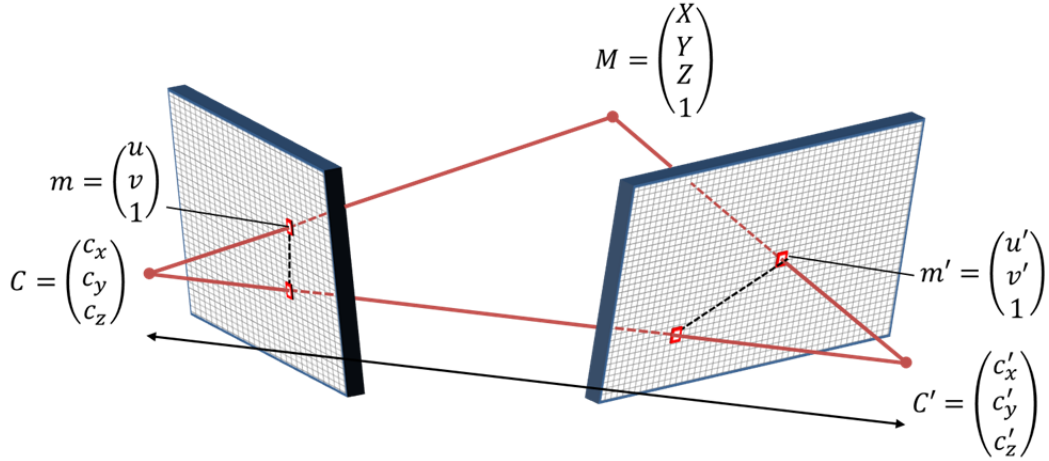


Figure 2.15. Stereo camera geometry.

Figure 2.15 illustrates the projection of a 3D world point $\mathbf{M} = (X, Y, Z, 1)$ onto two images planes with the respective pixel coordinates $\mathbf{m} = (u, v, 1)$ and $\mathbf{m}' = (u', v', 1)$. To calculate the projection process numerically, it is sufficient to know the respective projection matrices \mathbf{P} and \mathbf{P}' as introduced in the section above (see eqn. (2.21)) and the coordinates of the 3D world point.

2.4.2.1 Epipolar Geometry and the Fundamental Matrix

In many applications such as feature point matching and disparity estimation, the 3D world point coordinates are not known. In contrast, a feature of interest such as a corner might be visible in the first camera, e.g. at pixel position \mathbf{m} and the task is now to retrieve the corresponding pixel position \mathbf{m}' in the second camera. The question is raised if there are any constraints on the pixel position \mathbf{m}' given that the pixel coordinates \mathbf{m} and the relative position, the orientation and the intrinsic parameters of the two cameras are known. In fact, such a requirement exists and is usually denoted as epipolar constraint [Hartley04]. The qualitative geometric derivation is the following: As shown in Figure 2.15, one can define a plane π in 3D space using the light ray or line which passes through \mathbf{m} and \mathbf{C} on one hand, and the line connecting the two camera centers \mathbf{C} and \mathbf{C}' on the other hand. The former ray is emitted by \mathbf{M} whose position in 3D space might be unknown a priori. Nevertheless, the projection of \mathbf{M} onto the imaging plane of the second camera needs to lie in the plane π , i.e. \mathbf{m}' lies on the line defined by the intersection of π and the imaging plane. This line is called epipolar line, while the projection of the camera centers \mathbf{C} and \mathbf{C}' onto the image planes are denoted as epipoles. The geometry describing the relationship between corresponding light rays as shown in Figure 2.15 is called epipolar geometry. The epipolar constraint can be expressed mathematically as follows [Hartley04]:

$$\mathbf{m}'^T \cdot \mathbf{F} \cdot \mathbf{m} = 0, \quad (2.25)$$

where \mathbf{F} is called the fundamental matrix. The pixel coordinates \mathbf{m} and \mathbf{m}' need to fulfill this requirement in order to be a putative match. However, the equation above imposes a necessary but not

sufficient requirement for pixel correspondences. The fundamental matrix is a 3x3 matrix with rank 2 and is defined as follows [Hartley04]:

$$\mathbf{F} = \mathbf{K}'^{-T} [\mathbf{t}]_{\times} \mathbf{R}' \cdot \mathbf{K}^{-1}, \quad (2.26)$$

where \mathbf{K} and \mathbf{K}' are the intrinsic matrices from the first and second camera as defined in eqn. (2.19). The rotation matrix \mathbf{R}' describes the relative orientation of the second camera with respect to the first camera. The translation between the two camera centers is described by the term $[\mathbf{t}]_{\times}$ which is defined as follows:

$$[\mathbf{t}]_{\times} \stackrel{\text{def}}{=} \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}. \quad (2.27)$$

The translation vector \mathbf{t} is defined as:

$$\mathbf{t} = -\mathbf{R}' \cdot \mathbf{C}'. \quad (2.28)$$

In eqn. (2.28), \mathbf{C}' is the camera center of the right camera while the center of the first camera \mathbf{C} coincides without loss of generality with the origin.

To check how well two pixel-coordinates \mathbf{m} and \mathbf{m}' fulfill the epipolar constraint from eqn. (2.25) the Sampson distance [Hartley04] defined in the following equation can be calculated:

$$\sum_i \frac{(\mathbf{m}'^T \mathbf{F} \mathbf{m})^2}{(\mathbf{F} \mathbf{m})_1^2 + (\mathbf{F} \mathbf{m})_2^2 + (\mathbf{F}^T \mathbf{m}')_1^2 + (\mathbf{F}^T \mathbf{m}')_2^2} \quad (2.29)$$

where $(\mathbf{F} \mathbf{m})_i$ denotes the i^{th} element of the product of the 3x3 matrix \mathbf{F} and the 3x1 vector \mathbf{m} . The equivalent of the fundamental matrix without intrinsic camera parameters is the essential matrix \mathbf{E} [Faugeras93b] defined as follows:

$$\mathbf{E} = [\mathbf{t}]_{\times} \mathbf{R}'. \quad (2.30)$$

2.4.2.2 Stereo Rectification and Related Work

Within a two-camera rectification, a pair of homographies is searched which ensures that after applying the rectifying homographies, the optical axes of the two cameras as well as the two image planes are parallel. The basic concept is visualized in Figure 2.16.



Figure 2.16. Concept of the image rectification process. Both cameras are virtually rotated by applying a rectifying homography such that the new stereo baseline (dotted red line) is perpendicular to the two cameras' orientations while normalizing the intrinsic matrices.

Faugeras describes rectification as a reprojection of the left and the right image onto a common image plane [Faugeras93b]. By rectification he aims to ensure a simplified epipolar geometry, i.e. the epipoles are at infinity and the epipolar lines match with the image scanlines which facilitates dense stereo matching. The new image plane needs to be parallel to the baseline. In the rectified state, the two intrinsic matrices are identical, i.e. $\mathbf{K}' = \mathbf{K}$. Moreover, the orientations match, i.e. $\mathbf{R}' = \mathbf{I}$ and the translation between the two cameras is purely horizontal, i.e. $\mathbf{t} = (t_x, 0, 0)^T$. It is easy to show using eqn. (2.26) that in this case, the fundamental matrix has the form:

$$\mathbf{F}_{rect} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}. \quad (2.31)$$

By inserting \mathbf{F}_{rect} into the epipolar constraint from eqn. (2.25), one can show that for two corresponding pixel coordinates $\tilde{\mathbf{m}} = (\tilde{u}, \tilde{v}, 1)^T$ and $\tilde{\mathbf{m}}' = (\tilde{u}', \tilde{v}', 1)^T$, of the rectified image pair, the following equation holds true:

$$\tilde{v} = \tilde{v}'. \quad (2.32)$$

In other words, no vertical but only horizontal disparities are left. The epipolar lines are parallel and coincide with the image scanlines. Figure 2.17 illustrates the geometry of the rectified image pair.

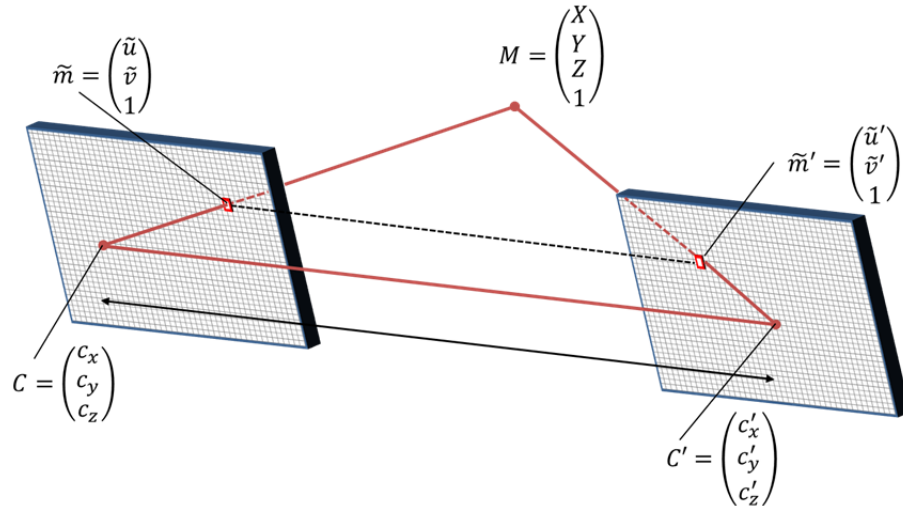


Figure 2.17. Rectified image pair. The epipolar lines are parallel, no vertical disparities are left.

The image planes have been rotated compared to Figure 2.15 but the positions of the camera centers \mathbf{C} and \mathbf{C}' have not been changed. The rotation and normalization of the intrinsic parameters has been performed by applying rectifying homographies \mathbf{H} and \mathbf{H}' to the two cameras camera projection matrices \mathbf{P} and \mathbf{P}' . As a result the two image planes are now parallel. It is the task of any stereo rectification algorithm to find a suitable common image plane and the associated pair of rectifying homographies \mathbf{H} and \mathbf{H}' . However, there are two degrees of freedom to choose such a plane. While one of the degree of freedom only affects a possible scaling of the images, the other parameter is responsible for image distortion effects. From Figure 2.17 one can see that one could rotate the two image planes around the line connecting the two camera centers while fulfilling the simplified epipolar constraint from eqn. (2.32).

Stereo Disparity

A 3D world point $\mathbf{M} = (X, Y, Z, 1)^T$ is projected to the two pixel positions $\mathbf{m} = (u, v, 1)^T$ in the left and $\mathbf{m}' = (u', v', 1)^T$ in the right camera. The difference of the pixel positions $\mathbf{m}' - \mathbf{m}$ is called disparity. In the case of a rectified stereo pair, only horizontal disparities are left. The disparity can be calculated as follows. Without loss of generality, it can be assumed that the center of origin coincides with the left camera center, i.e. $\mathbf{C} = (0, 0, 0)^T$ and that the orientation of the two cameras is chosen such that the center of the right camera is given by $\mathbf{C}' = (\mathcal{B}, 0, 0)^T$, where \mathcal{B} denotes the stereo baseline. Assuming that both cameras have identical intrinsic matrices and that the normal vector of the image planes coincides with the z-axis of the world coordinate system, the projection matrices \mathbf{P}_L and \mathbf{P}_R of the left and right cameras can be written according to eqn. (2.21) as follows:

$$\mathbf{P}_{left} = \begin{bmatrix} f & 0 & u_0 & 0 \\ 0 & f & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{P}_{right} = \begin{bmatrix} f & 0 & u_0 & -f \cdot \mathcal{B} \\ 0 & f & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \quad (2.33)$$

Using eqn. (2.22), the projected points \mathbf{m} and \mathbf{m}' can be calculated as follows:

$$\mathbf{m}' - \mathbf{m} = \mathbf{P}_{right} \cdot \mathbf{M} - \mathbf{P}_{left} \cdot \mathbf{M}, \quad (2.34)$$

which results in:

$$\mathbf{m}' - \mathbf{m} = \begin{pmatrix} fX + Z \cdot u_0 - f\mathcal{B} \\ Z \cdot v_0 \\ Z \end{pmatrix} - \begin{pmatrix} fX + Z \cdot u_0 \\ Z \cdot v_0 \\ Z \end{pmatrix} = \begin{pmatrix} f(X - \mathcal{B})/Z + u_0 \\ v_0 \\ 1 \end{pmatrix} - \begin{pmatrix} fX/Z + u_0 \\ v_0 \\ 1 \end{pmatrix}. \quad (2.35)$$

It is now possible to calculate the horizontal disparity d :

$$d = u' - u = f \cdot \frac{\mathcal{B}}{Z}. \quad (2.36)$$

Related Work

Image rectification is well known in literature. Faugeras proposes to choose the common image plane such that it is parallel to the line of intersection of the two original image planes [Faugeras93b]. Zhang et al. propose an algorithm which combines the estimation of the epipolar geometry with a guided point matching [Zhang95]. Papadimitriou and Dennis describe a vertical registration algorithm [Papadimitriou96]. Hartley proposes a rectification algorithm which is suitable for wide baseline systems [Hartley99]. A main idea of this algorithm is to minimize horizontal disparities in order to facilitate image matching. Furthermore, Hartley claims that the image center undergoes a rigid transformation, i.e. only rotation and translation are applied to the image center. Loop and Zhang propose a rectification algorithm which reduces image distortions by decomposing the rectifying homographies into a similarity transform, followed by a shearing transform [Loop99]. Isgrò and Trucco propose to calculate rectifying homographies without explicit knowledge of the epipolar geometry [Isgrò99]. Fusiello et al. propose a linear rectification algorithm based on two perspective projection matrices [Fusiello00]. Wu and Yu propose to minimize distortion by using a properly chosen shearing transform [Wu05]. Mallon and Whelan compare different rectification algorithms with respect to their image distortion impact [Mallon05]. They derive rectifying homographies from a fundamental matrix which might be affected by noise. In [Georgiev13], a rectification method without explicit estimation of the fundamental matrix for a stereo setup with constraint degrees of freedom is proposed.

In chapter 3, a stereo rectification algorithm which has improved alignment performance compared to the above mentioned methods and which is optimized for the production stereoscopic 3D content will be proposed.

2.4.3 Three-Camera Geometry

In this section, the extension from two-camera geometry to three-camera geometry is described. Figure 2.18 illustrates the geometrical setup. A 3D world point \mathbf{M} is projected onto three image planes from three cameras located at their respective camera centers \mathbf{C} , \mathbf{C}' , and \mathbf{C}'' . Using the projection matrices \mathbf{P} , \mathbf{P}' and \mathbf{P}'' parameterizing the first, second and third camera, it is possible to calculate the pixel coordinates \mathbf{m} , \mathbf{m}' and \mathbf{m}'' of the intersection of the light rays emitted from \mathbf{M} which pass through the different camera centers.

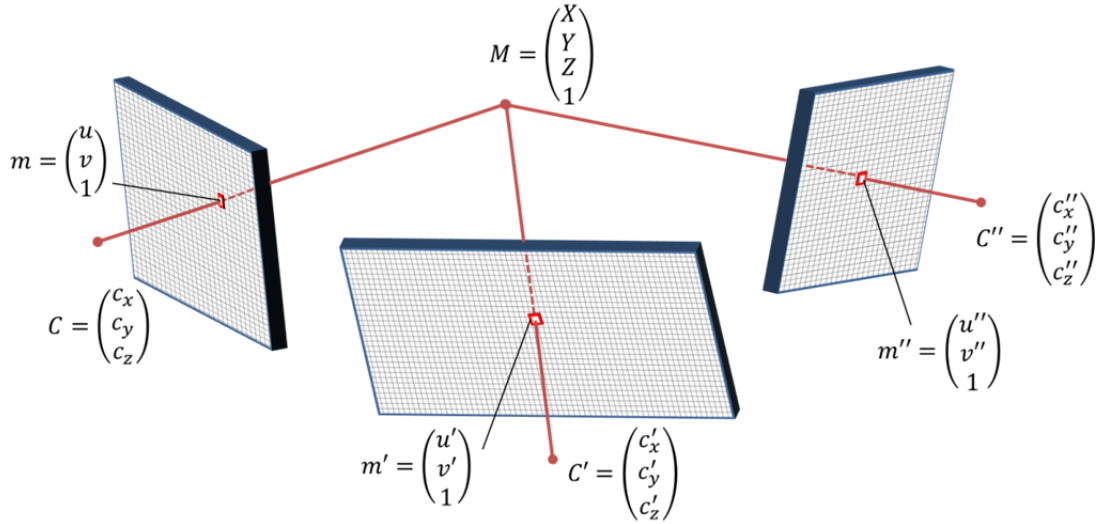


Figure 2.18. Three-camera geometry.

An interesting aspect of the three-camera geometry is that once all the projection matrices of the three cameras are known, one can conclude from the pixel coordinates \mathbf{m} and \mathbf{m}' the possibly missing coordinate \mathbf{m}'' . The relation between the three cameras is described by the trifocal tensor, which plays a similar role as the fundamental matrix in the two-camera geometry.

In the following, a method presented in [Hartley04] to calculate the trifocal tensor given three projection matrices \mathbf{P} , \mathbf{P}' and \mathbf{P}'' will be presented. Thereby it is assumed that the camera center \mathbf{C} of the first camera coincides with the origin and that the optical axis coincides with the z-axis of the world coordinate system. Moreover, it is assumed that the intrinsic parameters are normalized such that the projection matrix \mathbf{P} has the following form:

$$\mathbf{P} = [I|0]. \quad (2.37)$$

By representing the projection matrices \mathbf{P}' and \mathbf{P}'' of the second and third camera by their components

$$\mathbf{P}' = [a_i^j], \mathbf{P}'' = [b_i^j], \quad (2.38)$$

where a_i^j and b_i^j represent the elements in the i^{th} column and j^{th} row of the projection matrix \mathbf{P}' and \mathbf{P}'' respectively, the trifocal tensor can be calculated as follows [Hartley04]:

$$\mathcal{T}_i^{jk} = a_i^j b_4^k - a_4^j b_i^k. \quad (2.39)$$

The trifocal tensor consists of 27 elements. It can also be represented in slices, i.e. by three matrices \mathcal{T}_1^{jk} , \mathcal{T}_2^{jk} and \mathcal{T}_3^{jk} with 3×3 elements each.

Given trifocal point correspondences $\mathbf{m} = (u, v, 1)^T$, $\mathbf{m}' = (u', v', 1)^T$ and $\mathbf{m}'' = (u'', v'', 1)^T$ from three cameras, it is possible to perform a consistency check of the pixel coordinates using the following relation using the cross-product notation from eqn. (2.27):

$$\begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix}_{\times} (u\mathcal{T}_1 + v\mathcal{T}_2 + \mathcal{T}_3) \begin{bmatrix} u'' \\ v'' \\ 1 \end{bmatrix}_{\times} = \mathbf{0}_{3 \times 3}. \quad (2.40)$$

Each point triplet gives rise to nine equations wherefrom four are linear independent [Hartley04, Shashua95]. For a linear camera array where all camera centers lie on a common baseline, the above formula transforms to:

$$u'' = \frac{u(\mathcal{T}_1^{11} - u'\mathcal{T}_1^{31}) + v(\mathcal{T}_2^{11} - u'\mathcal{T}_2^{31}) + (\mathcal{T}_3^{11} - u'\mathcal{T}_3^{31})}{u(\mathcal{T}_1^{13} - u'\mathcal{T}_1^{33}) + v(\mathcal{T}_2^{13} - u'\mathcal{T}_2^{33}) + (\mathcal{T}_3^{13} - u'\mathcal{T}_3^{33})}. \quad (2.41)$$

The vertical components need to fulfill the following equation in order to be trifocal consistent:

$$v'' = \frac{u(\mathcal{T}_1^{12} - u'\mathcal{T}_1^{32}) + v(\mathcal{T}_2^{12} - u'\mathcal{T}_2^{32}) + (\mathcal{T}_3^{12} - u'\mathcal{T}_3^{32})}{u(\mathcal{T}_1^{13} - u'\mathcal{T}_1^{33}) + v(\mathcal{T}_2^{13} - u'\mathcal{T}_2^{33}) + (\mathcal{T}_3^{13} - u'\mathcal{T}_3^{33})}. \quad (2.42)$$

2.4.3.1 Multi-Camera Rectification and Related Work

The trifocal tensor and the relation of three point correspondences presented in eqn. (2.40) can be used to check if three pixel coordinates correspond to the same 3D world point. However, in order to create pixel dense depth maps using three cameras, the question is raised if a simplified consistency check is possible, provided that the cameras are arranged in a special geometry and that the intrinsic parameters are normalized similarly to the stereo rectification in the two-camera case. In this context, two different camera setups are of particular interest, the linear camera array where all three camera centers are placed on a common baseline and the L-shaped setup where the baselines between first and second camera on hand and the baseline between second and third camera form an angle of 90 degrees. Both setups allow a simple computation of pixel correspondences after normalization or rectification of the cameras.

Related Work

Several trifocal or trinocular rectification algorithms are known in the literature which will be presented in the following. Trinocular rectification techniques which require calibrated cameras or dedicated calibration pattern have been proposed in [An04], [Ayache88], [Boutarel10], [Xin04] and [Kang08]. In contrast, the method proposed by [Sun03] works with uncalibrated cameras by using the trilinear tensor in its representation used in [Shashua95]. Other calibration-free approaches for trifocal rectification were proposed in [Heinrichs06], [Baik07] and [Zhang03]. However, these techniques

target L-shaped trifocal setups while within this thesis, the focus lies on linear cameras arrays. A calibration for multi-view panoramic camera-setups was recently proposed by Kurillo et al. [Kurillo13]. A rectification method for three cameras in a horizontal setup using uncalibrated cameras has been proposed by [Kangni06] and extended towards four or more cameras by [Yang10]. Both methods are based on [Mallon05]. In 2011, Nozick extended the approach from Boutarel and Nozick [Boutarel10] towards uncalibrated cameras [Nozick11]. A simple, though real-time capable multi-camera rectification algorithm which considers only vertical pixel shifting has been proposed in [He10]. A simple horizontal pixel shifting to achieve equidistant disparities is considered in [Yang10]. However, parallel epipolar lines are not a sufficient constraint for a proper horizontal alignment, as a stretching or offset of the horizontal disparities can still occur due to a horizontal shift of the principal point or a deviation of the pixel aspect ratio from 1. In chapter 3, a method will be presented which takes into account the horizontal alignment in contrast to the previously mentioned approaches.

2.5 Feature Point Matching

2.5.1 Introductory Remarks

Feature Point detection and matching is a vast field of research for its own and has dozens of applications in computer vision such as object tracking and recognition, camera pose estimation and self-calibration¹⁰. In the context of this thesis, reliable point correspondences along two or more cameras will be used for calibration purposes and to measure the amount of depth within a scene.

In this chapter, an overview of the related work and the basic concepts of feature detection and matching is given. Thereby, the structure of this section is aligned to the three basic steps of feature detection and matching as shown in Figure 2.19, namely interest point detection, interest point description and finally, matching of the formerly described key points.

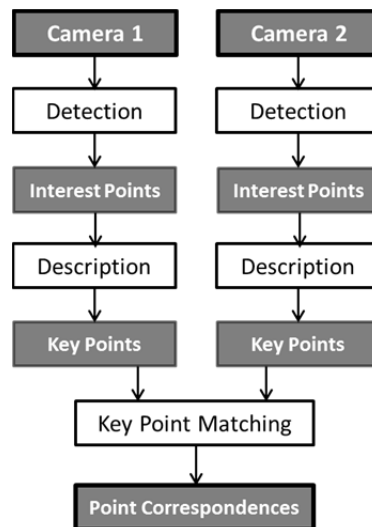


Figure 2.19. Main steps of a feature point based algorithm aiming to establish point correspondences within a stereo image pair.

¹⁰ Parts of the content in this section have been previously published in [Zilly11c].

As already mentioned, many application for feature detection and matching exist and the concepts described in this chapter can be applied to most of these applications. However, the case of matching point correspondences within a stereo image pair as shown in Figure 2.19 is of particular interest for this thesis. Related applications will be presented in chapters 4 and 5. According to these applications, the focus of this chapter will be feature detectors based on blobs rather than corners.

2.5.2 Interest Point Detection and Related Work

The first step within a feature detection process is to find interest points within an image. An interest point can be a visible corner in the image, or the center of a blob, or any other feature e.g. based on saliency criteria which are suitable to distinguish a particular point or region from others in the image. In [Mikolajczyk05b] the term region detector is used to underlie the fact that not only a single pixel or point but a whole region or set of pixels might be of interest. The specific image properties such as cornerness, blobness, or saliency are usually expressed by extremal values of a defined function in scale space such as Laplacian, Difference of Gaussian, Determinant of Hessian functions, or other structure tensor functions. Some region detectors are even invariant against affine transformations [Kadir04, Mikolajczyk04, Mikolajczyk05b].

In the past years much attention has been paid to the development of robust interest point detectors and descriptors. SIFT [Lowe04] and SURF [Bay08] are two prominent examples of such combined interested point detectors with associated descriptors. Moreover a number of dedicated region detectors have been proposed [Lindeberg98, Matas04, Tuytelaars04, Harris88, Ebrahimi09]. In-depth comparisons of the different detectors and descriptors have been carried out in [Mikolajczyk05a] and [Mikolajczyk05b].

2.5.2.1 Scale Space Pyramid

The search for interest points can be carried out in scale space, in order to identify not only the position of an image feature, but also the scale where the image feature has the highest discriminative power. To allow a search in scale space as performed in [Lowe04], the images are subsequently subsampled or filtered using Gaussian filters of growing standard deviation. The images stored in a scale space pyramid are then folded with filter kernels which are suitable to identify image features. The filter responses are then stored in a filter response pyramid. The pyramid consists of n octaves while each octave consists of m intervals, i.e. m different scales exist per octave. Within the first octave, all pixels are evaluated. For the subsequent octave, a 1:2 subsampling in x - and y -direction is applied [Lowe04], in the next octave a 1:4 subsampling etc. This saves processing time and memory usage for the filter response images. The concept of neighboring scales and pixels is illustrated in Figure 2.20.

Once the initial interest point has been detected, a region around this interest point, nominated as support region, is extracted from the image. The support region might be normalized in scale and

possibly in orientation to achieve scale invariance and rotational invariance respectively. An automatic scale selection is commonly performed as proposed in [Lindeberg98].

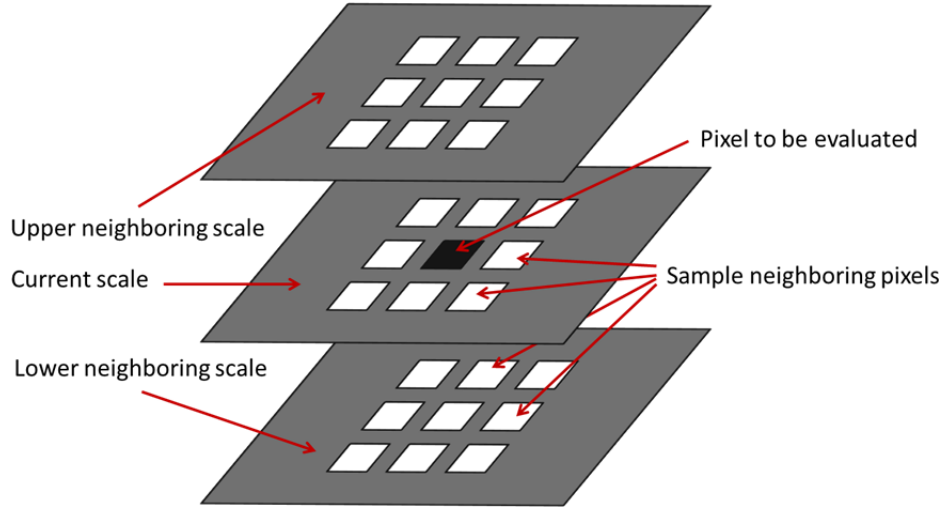


Figure 2.20. Concept of neighboring scales. Each pixel has eight neighbors in the current scale and nine neighbors in the upper and lower neighboring scale respectively which results in a total of 26 neighboring pixels. The non-maximum suppression and local search for extremal values is performed in this neighborhood.

2.5.2.2 Blob Detectors

Many state-of-the-art algorithms use a blob detector to detect interest points in scale space. Blobs are regions whose grayscale value near the center is higher (or lower) than the surrounding pixels. The Laplacian of Gaussian function or similar functions such as the Mexican hat function can be used to detect blobs. As the function needs to be evaluated for all pixels in all different scales, one tries to find efficient implementations or approximations in order to achieve reasonable processing speed. In this context, [Lowe99] showed that a Difference of Gaussians can be used to approximate the Laplacian of Gaussian. In a first step, images within the same octave in the scale space pyramid are folded with Gaussian kernels with different standard deviations σ as defined in eqn. (2.43):

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \quad (2.43)$$

Subsequently, the pixel-wise difference of these blurred images, e.g. with the standard deviations σ and $k\sigma$ is calculated. Eqn. (2.44) shows that the result approximates the Laplacian of Gaussian function $\nabla^2 G$ which is suitable to detect blobs:

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k - 1)\sigma^2 \nabla^2 G. \quad (2.44)$$

The SURF interest point detector [Bay08] tries to find blobs by searching for local extrema of the determinant of the Hessian function denoted as fast hessian detector. The Hessian matrix from eqn. (2.45)

$$\mathbf{H}(x, y, \sigma) = \begin{bmatrix} L_{xx}(x, y, \sigma) & L_{xy}(x, y, \sigma) \\ L_{xy}(x, y, \sigma) & L_{yy}(x, y, \sigma) \end{bmatrix} \quad (2.45)$$

is composed of the second order partial derivatives $L_{xx}(x, y, \sigma)$, $L_{yy}(x, y, \sigma)$ and $L_{xy}(x, y, \sigma)$. The former is defined as:

$$L_{xx}(x, y, \sigma) = \frac{\partial^2}{\partial x^2} G(x, y, \sigma). \quad (2.46)$$

The remaining partial derivatives are defined accordingly. The evaluation of the components of the Hessian matrix is performed using box filters and integral images, a concept which will be explained in the following.

2.5.2.3 Box Filters and Integral Images

Integral images, also known as summed area tables and were introduced by [Viola01] for the fast evaluation of box filters. It is assumed that $\mathbf{Img}(x, y)$ denotes the intensity of the pixel in the column x and the row y and that the origin of the image is in the top left corner and has the coordinates $x_{origin} = 0$ and $y_{origin} = 0$. The Integral Image at this pixel position $\mathbf{IntImg}(x, y)$ is then defined as follows:

$$\mathbf{IntImg}(x, y) = \sum_{x_i=0}^x \sum_{y_i=0}^y \mathbf{Img}(x_i, y_i). \quad (2.47)$$

Once the integral image has been calculated, it can be used to evaluate the integral of rectangular surfaces within the image. This concept is illustrated in Figure 2.21. According to eqn. (2.47), the summed grayscale in area **I** of the original image can be fetched by evaluating the integral image at pixel position **a** with the coordinates (x_a, y_a) , i.e. $\mathbf{IntImg}(x_a, y_a) = \sum \mathbf{I}$. Accordingly, the pixel position **b** of the integral image relates to the conjunction of the surfaces **I** and **II**, i.e. $\mathbf{IntImg}(x_b, y_b) = \sum \mathbf{I} + \sum \mathbf{II}$ while $\mathbf{IntImg}(x_d, y_d) = \sum \mathbf{I} + \sum \mathbf{III}$ and $\mathbf{IntImg}(x_c, y_c) = \sum \mathbf{I} + \sum \mathbf{II} + \sum \mathbf{III} + \sum \mathbf{IV}$.

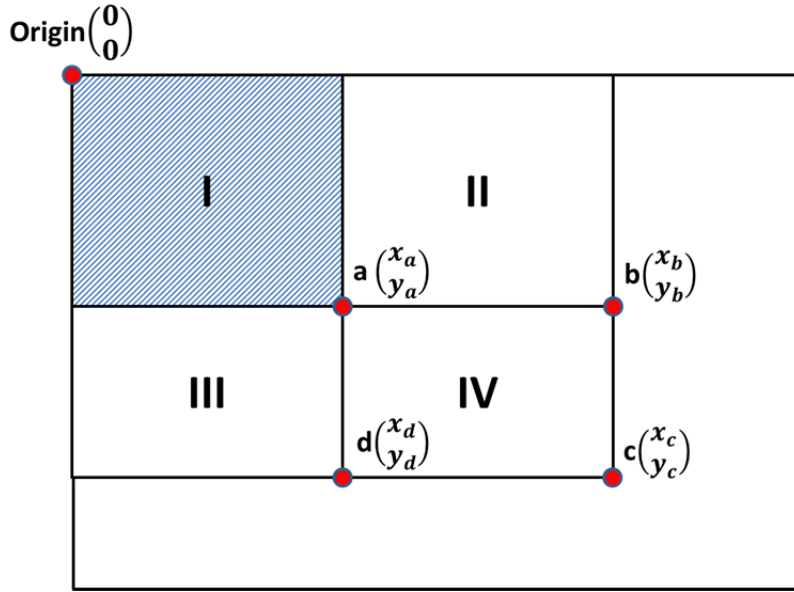


Figure 2.21. Concept of integral images. The grayscale value of each pixel in the integral image corresponds to the sum of all pixels from the input image which lie in the rectangle defined by the pixel position and the origin. The pixel sum in the hatched surface *I* can be fetched by reading the pixel $a(x_a, y_a)^T$ from the integral image.

To evaluate for instance the pixel-sum of area *IV*, the following algebraic operation can be performed:

$$IntImg(x_a, y_a) + IntImg(x_c, y_c) - IntImg(x_b, y_b) - IntImg(x_d, y_d) = \sum IV \quad (2.48)$$

or, in order to shorten the expression and the respective nomenclature:

$$a + c - b - d = \sum IV. \quad (2.49)$$

The above described concept is used within the SURF feature detector [Bay08] and the evaluation of the Hessian matrix from eqn. (2.45). For an efficient evaluation of the second order derivatives, the discretized and simplified filter kernels shown in Figure 2.22 are used.

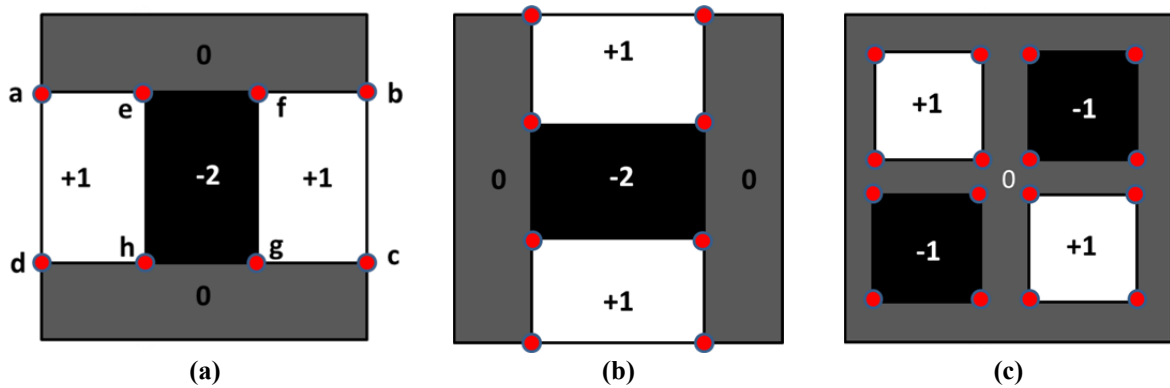


Figure 2.22. Box filter used to evaluate the Hessian of Gaussian. (a) Approximation D_{xx} for L_{xx} from eqns. (2.45) and (2.46). (b) Approximation D_{yy} for L_{yy} . (c) Approximation D_{xy} for L_{xy} . The red dots indicate the positions where a pixel fetch from the integral image is performed.

The filter kernel shown in Figure 2.22 (a) approximates the element L_{xx} from eqns. (2.45) and (2.46). As shown, its evaluation requires only eight pixel fetches from the integral image and has constant time irrespective of the scale. The filter response D_{xx} of the approximated second order derivative L_{xx} can be evaluated as follows:

$$D_{xx} = a + c - b - d - 3 \cdot (e + g - f - h). \quad (2.50)$$

The filter kernels from Figure 2.22 (b) and (c) can be evaluated using the same approach. The red dots indicate the positions where a pixel fetch from the integral image is performed. In a last step, the determinant of the approximated Hessian matrix is calculated. According to [Bay08], the following formula which inhibits a correction term for a deviation of the box filters from ideal Gaussian-based kernels leads to the best results:

$$\det(H_{approx})(x, y, \sigma) = D_{xx}D_{yy} - (0.9 D_{xy})^2. \quad (2.51)$$

The approach followed by the SUSurE feature detector [Ebrahimi09] is based on an even further going simplification. A single binarized kernel is used to detect extremal values of the Laplacian of Gaussian function. The detector uses only a single filter operation performed on an integral image in the box filter variant of SUSurE. The kernel is particularly efficient to evaluate. Figure 2.23 shows the kernel used for the convolution. The innermost square has positive weights, surrounded by a square with negative weights. The black and the white region have the same surface. The outer region (grey) does not need to be considered and the respective weights have been set to zero.

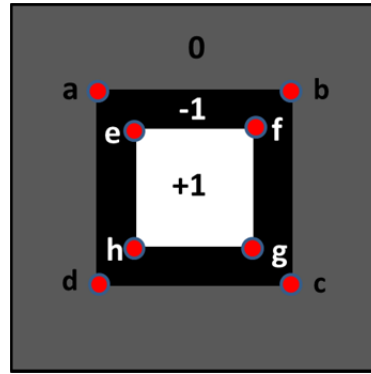


Figure 2.23. Kernel used for searching extremal values of the Laplacian of Gaussian in scale space.

The filter response r of the SUSurE blob filter requires only eight pixels fetches and can be evaluated as follows:

$$r = (e + g + d + b) - (a + c + f + h). \quad (2.52)$$

Similarly to the SURF feature detector kernels from Figure 2.22, the convolution operation is constant in time over all scales.

2.5.2.4 Non-Maximum Suppression and Search for Local Extrema

After the filter responses have been evaluated, one searches for local extrema in the filter response images. Commonly, all pixels in the convoluted images above a threshold t are examined. Therefore, a neighborhood of 26 surrounding pixels is evaluated – 8 neighbors in the current scale, and 2x9 neighbors in the adjacent scales as shown in Figure 2.20. The search for local extrema can be efficiently implemented as proposed in [Neubeck06].

2.5.2.5 Refined Interest Point Localization

The local maxima which are above a threshold t are candidates for interest points. In the higher octaves, the sampling of the filter responses is coarser than in the lower octaves, with respect to the integral image. Therefore, an additional sampling can be performed for the filter response around candidate pixels in higher octaves by interpolating the subpixel location of the interest point in scale and position. A suitable mechanism is to build a Taylor polynomial of second degree as proposed by [Brown02]. The interest point candidates are not numerically stable if they do not correspond to a well-defined blob, but to a ridge-like structure which might also yield to filter responses above the threshold. A usual approach is to discard the interest point candidate if the interpolated subpixel position has a distance of more than $\frac{1}{2}$ pixel from the original integer pixel position.

Another approach is to explicitly reject edges as performed within the DoG search used by SIFT [Lowe04].

2.5.2.6 Extraction of the Support Region

After the subpixel interpolation, the interest point has an assigned scale and position, which can later be used by the descriptor to define a suitable support region. To achieve rotational invariance, the main gradient direction can be extracted in order to *turn* the interest point (and later the support region) such that the direction points northwards. However, for some applications such as stereo matching with nearly rectified cameras, rotational invariance is not necessary. In this case a rotationally variant detector such as U-SURF [Bay08] can be used.

2.5.3 Interest Point Description and Related Work

The commonly used descriptors, such as SIFT [Lowe04], SURF [Bay08], or GLOH [Mikolajczyk05a] use histograms of gradients to describe the support region. High dimensional feature vectors are used to describe the feature point. SURF uses 64 dimensions while SIFT uses 128 dimensions. Each dimension is typically represented by a floating point number or by a 1-byte integer value. This leads to a high memory usage and a fairly slow matching process.

Consequently, effort has been spent to reduce the dimensionality. PCA-SIFT proposed by [Ke04] and GLOH [Mikolajczyk05a] reduce the dimensions by performing a principal component analysis of the feature vectors. However, the description process is slowed down by this operation. Moreover, the

descriptors itself remain floating point numbers or integer values which use the relatively slow L2-norm to match. On the other hand, the matching quality in terms of recall rate is much higher than simpler interest point detectors such as the Harris Corner detector [Harris88] in combination with a cross correlation based matching.

In the recent past, research has been conducted to build descriptors consisting of binary elements, denoted as binary strings in order to reduce the bitrate needed to transmit and to store the descriptors and to reduce the matching time, as matching can then be performed using the Hamming distance which can be efficiently computed on modern CPUs. Different descriptor binarization strategies exist. Some approaches first compute a higher dimensional descriptor, which will later be binarized by hashing the real-valued interest point descriptors [Ventura11], by performing a transform coding [Brandt10] or by direct binarization of the descriptor components [Stommel10]. On the other hand, an even further speed up can be achieved when the description process leads directly to a binarized descriptor. A prominent example for this method is the BRIEF descriptor [Calonder10]. A specific number of intensity comparisons (e.g. 256 in the case of BRIEF-256) is conducted which results in a 256-bit descriptor. A similar concept is followed by the descriptor SKB [Zilly11c] which is presented in chapter 4. During the description process, 16 different kernels are evaluated at 16 positions in the support region resulting in 256 bit comparisons. A hardware implementation of the SKB descriptor was proposed in [Schaffner13].

2.5.4 Interest Point Matching

In order to establish point correspondences between feature points from different images, a feature point matching needs to be applied which compares two or more feature vectors and measures a similarity or distance resulting in a matching score. Thereby, matching two feature vectors is usually much faster and more robust than measuring the distance between two image patches, e.g. by using a cross correlation or SAD. Nevertheless, the runtime of the matching process remains an issue, especially if a high number of putative matches need to be evaluated. As mentioned in the previous section, the length of the feature vector (e.g. 128 dimensions for SIFT) and the bit-depth per dimension influences the matching speed. The matching process itself can be divided into three main steps. First, the similarity or distance between two vectors is evaluated e.g. using the scalar product or Euclidean distance. Subsequently, additional matching constraints, such as uniqueness of the matching pair can be applied. Finally, during a post-processing step adapted to the application, further constraints such as the epipolar constraint from section 2.4.2.1 can be applied to the list of putative matches. In the following, the above mentioned processing steps will be presented.

2.5.4.1 Distance Metrics

If the feature vector consists of floating point numbers, it can easily be normalized to unit length [Lowe04]. Subsequently, the scalar product s or inner product can be calculated as usual, e.g. by multiplying the components of the n -dimensional feature vector and summing up all products:

$$s = \sum_{i=1}^n a_i \cdot b_i \quad (2.53)$$

To evaluate the binary scalar product s , a binary **AND** is performed between the feature vectors a and b . The concept is visualized in Figure 2.24. Subsequently, the number of bits set is evaluated, namely the population count as described in eqn. (2.54).

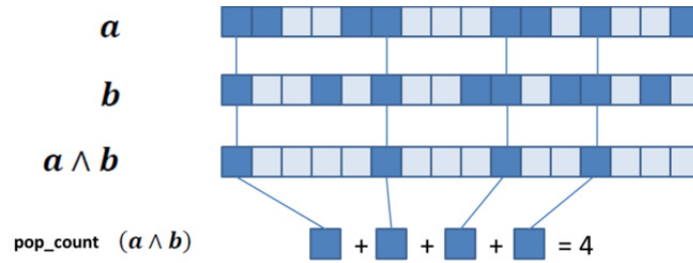


Figure 2.24. Binarized scalar product with subsequent evaluation of the population count.

$$s = \text{pop_count}(a \wedge b). \quad (2.54)$$

The above operation can be performed very efficiently on modern CPUs. For instance, when using 64 bit registers, only four CPU operations are required to calculate the bit-wise **AND** between two feature vectors of 256 bit length. Subsequently, the number of bits set is evaluated using the population count. This process is illustrated in Figure 2.24. Modern CPUs support SSE4.2 [Intel07] which has dedicated instructions to calculate the population count. The CPU instruction `pop_cnt64` evaluates the population count of a 64 bit number.

Another distance metric suitable for binarized feature descriptors such as BRIEF [Calonder10] is the Hamming distance. The Hamming distance h between two binarized feature vectors a and b is calculated using the binary **XOR** and subsequent evaluation of the population count:

$$h = \text{pop_count}(a \vee b). \quad (2.55)$$

The Hamming distance can be thought as the Manhattan distance of two feature vectors on an n -dimensional unit-cube.

2.5.4.2 Additional Matching Constraints

In the simplest case, the feature vector with the highest evaluated matching score is identified as match. However, additional matching constraints can be applied which help to increase the robustness of the feature matcher. An obvious approach is to reject matches, if the matching score is below a threshold θ_{score} . Another common approach is the so-called uniqueness constraint. Literally speaking, a putative match is rejected if the matching score of the best match $s1$ is not significantly better than the score of the second best match $s2$, i.e. if the ratio between $s1$ and $s2$ is lower than a dedicated threshold $\theta_{uniqueness}$, i.e. if $\frac{s1}{s2} < \theta_{uniqueness}$.

2.5.4.3 Performance Metrics

The quality of the feature point matching process can be measured using the recall rate and the value “1-precision”. According to [Ke04] the recall is defined as:

$$recall = \frac{\text{number of "correct-positives"}}{\text{total number of positives}}, \quad (2.56)$$

where the value “1-precision” is the outlier rate:

$$1 - precision = \frac{\text{number of "false-positives"}}{\text{total number of matches (false or positive)}}. \quad (2.57)$$

The notation from [Ke04] will be used for the evaluation of the feature descriptor SKB presented in chapter 4.

2.6 Disparity Estimation

The aim of all stereo disparity algorithms is to establish point correspondences on pixel level in contrast to feature detection and matching approaches described in section 2.5. Thereby, different disparities belonging to different putative pixel matches are usually evaluated following a set of different constraints, such as epipolar constraint, cross correlation of neighboring pixels, difference of the pixel intensities, smoothness of the disparity map and other criteria. If required, disparity estimation can be seen as minimizing an energy function whereas an energy term can be associated to different matching criteria.

2.6.1 Related Work

Stereo disparity estimation is a vital field of research since many years¹¹. Hundreds of different approaches are known in literature and new algorithms are proposed every year. An extensive overview of different disparity estimation techniques along with a proposed performance evaluation

¹¹ Parts of the content in this section have been previously published in [Zilly14].

scheme for different algorithms is given in [Scharstein02]¹². A disparity estimation algorithm based on graph cuts was proposed in [Boykov01] and [Bleyer07]. An algorithm based on dynamic programming with additional uniqueness constraint on pixel basis was proposed in [Cox96]. An algorithm suitable for real-time operation based on correlation techniques is described in [Faugeras93a]. A disparity estimation framework dedicated to real-time video conferencing based on visual hull shape estimation was proposed in [Feldmann09b]. An algorithm based on patch sweeping for a similar target application was proposed in [Waizenegger11].

In [Brown03], different correlation methods such as cross correlation, census or sum of absolute differences (SAD) are compared. Moreover, different strategies to cope with typical challenges within the matching process such as occlusions are evaluated. An in-depth inspection of the underlying geometry of half-occluded regions which gives insights for applications such as disparity estimation but also for a better understanding of the human visual system is performed in [Belhumeur96]. An algorithm dedicated to the precise estimation of depth discontinuities is proposed in [Birchfield98b]. An algorithm addressing the challenge of pixel sampling artifacts is proposed in [Birchfield98a].

In the simplest case, i.e. the stereo camera case, point correspondences between two views are searched. If the stereo pair is rectified, the search for corresponding pixels is 1-dimensional, as the corresponding pixels lie in the same image line. The concept of stereo disparity estimation can easily be extended to a search for correspondences among three or more views. If all views lie on a common baseline and are jointly rectified, the requirements of a multi-baseline scenario which allows evaluating the matching costs among more views as proposed in [Okutami93] are met. Thereby, the consistency between the disparities among different views can be evaluated once the different baselines are known. However, even in the multi-baseline scenario, according to Hirschmüller [Hirschmüller08] it can be preferable to transfer disparities between neighboring stereo pairs as occlusions in one or more views can disturb the overall correlation measure. In the case of three non-rectified cameras, one can check the consistency between point correspondences using a trifocal point transfer as presented in 2.4.3. With more than three cameras, one usually tries to map the pixel coordinates of the putatively corresponding pixel into 3D coordinates and back-projects them to the remaining cameras. The back-projection error, i.e. the average difference between the current and ideal pixel coordinates can thereby be used as consistency criterion. If the amount of available views is very high, an epipolar plane analysis is possible [Bolles87]. A recent improvement of this technique has been proposed by Kim et al. [Kim13].

2.6.2 Stereo Disparity Estimation

In the following, a typical workflow for stereo disparity estimation as shown in Figure 2.25 will be explained.

¹² The accompanying website is: <http://vision.middlebury.edu/stereo/>

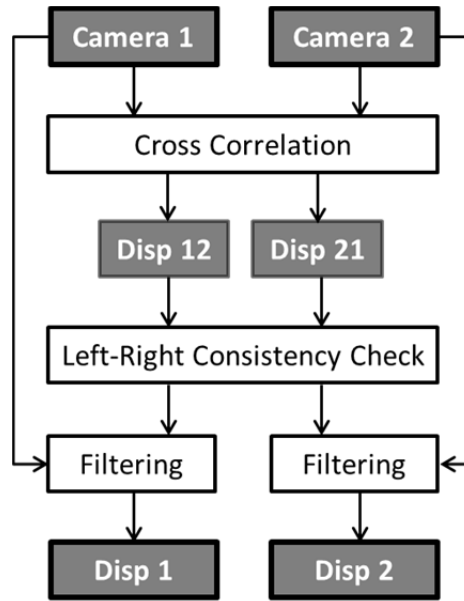


Figure 2.25. Stereo disparity estimation workflow as proposed in [Riechert12]. The generation of disparity maps is typically broken into different steps such as the similarity analysis (e.g. using a cross correlation), consistency checks (e.g. left-right consistency check), and post-processing e.g. by filtering of the disparity maps using morphological filters.

Given two rectified stereo images from camera 1 and camera 2, the similarity between different horizontally shifted pixels blocks, i.e. rectangular image patches from the two cameras is analyzed. This can be done using pixel metrics such as cross correlation, absolute difference of the intensity values, census transform or similar measures. In the easiest case, the pixel block with the highest similarity and the respective disparity is chosen. After processing all pixels two disparity maps result: one disparity map estimated from left to right and one which has been estimated from right to left. Subsequently, the consistency between these disparity maps can be checked. For instance, no valid disparities can be associated to those parts of the scene which have been visible only in one of the two cameras due to occlusions. The consistency check is able to reject these unreliable disparities. Finally, a post-filtering using cross-bilateral filters or morphological filters such as the median filter can be applied to get a smoother and more reliable disparity map.

In the following the mentioned processing steps will be described in more detail. The workflow can be seen as reference workflow to the multi-baseline disparity estimation workflow described in chapter 6.

2.6.3 Similarity Analysis

In the workflow illustrated in Figure 2.25, the similarity between different pixel blocks is compared. Thereby, one pixel block f is a cropped sub-image from the left view, while the second pixel block g originates from the right view. If the stereo pair has been rectified (see Figure 2.17 for an illustration), both pixel blocks were cropped around the same vertical position v as corresponding image features are expected to be seen in the same image scanline. However, the horizontal positions differ, i.e. f

might have been cropped at the pixel coordinate (u, v) from the left image, while \mathbf{g} might have been cropped at the pixel coordinate (u', v) from the right image yielding to a horizontal disparity d with:

$$d = u' - u. \quad (2.58)$$

A derivation of the above formula was performed in section 2.4.2.2.

The comparison between the two pixel blocks \mathbf{f} and \mathbf{g} can be calculated using the normalized cross correlation **ncc** which can be efficiently implemented according to [Lewis95] using the following formula:

$$\mathbf{ncc}(\mathbf{f}, \mathbf{g}) = \frac{\sum_{i,j} (f(i,j) - \bar{f}) \cdot (g(i,j) - \bar{g})}{\sqrt{\sum_{i,j} (f(i,j) - \bar{f})^2} \cdot \sqrt{\sum_{i,j} (g(i,j) - \bar{g})^2}} \quad (2.59)$$

where \bar{f} and \bar{g} denote the mean of the intensities within the blocks \mathbf{f} and \mathbf{g} respectively. Other similarity measures exist such as the *sum of absolute differences* **SAD** where the difference of the pixel intensities is calculated according to the following formula:

$$\mathbf{SAD}(\mathbf{f}, \mathbf{g}) = \sum_{i,j} |f(i,j) - g(i,j)|. \quad (2.60)$$

Yet another approach for a similarity measure is to apply the census transform to the pixel blocks \mathbf{f} and \mathbf{g} and then to evaluate the Hamming distance of the two resulting bit strings. Within the census transform, all pixel intensities within the block are compared to the center pixel. If the pixel under evaluation is brighter (or darker), the corresponding bit in the bit string is set to 1, or 0 in the inverse case. The concept of the census transform is also applied in a slightly modified way to the feature detector BRIEF [Calonder10].

2.6.4 Left-Right Consistency Check

According to the workflow illustrated in Figure 2.25, two disparity maps $Disp_{12}$ and $Disp_{21}$ result from the similarity analysis. The first disparity map $Disp_{12}$, contains disparities which point from a pixel position in the left image to a pixel position in the right image, while the inverse is true for $Disp_{21}$. Consistent disparities should thereby point to each other in the neighboring disparity maps. This is the basic idea of the left-right consistency which is performed according to eqn. (2.61) after the initial disparity estimation, e.g. the similarity analysis. A disparity is voted to be inconsistent if in the complementary disparity map the value does not point back to the initial position, i.e. if the target position lies more than the left-right consistency threshold $\theta_{l/r}$ away from the ideal position. Useful values for $\theta_{l/r}$ depend on the image resolution.

$$\forall i \in [1, \text{size}(\text{Disp}_{12})] : d_{12} = \text{Disp}_{12}(i),$$

$$\text{Disp}_{12}(i) = \begin{cases} d_{12} & \text{if } |\text{Disp}_{21}(i + d_{12}) + d_{12}| < \theta_{l/r} \\ \text{invalid} & \text{else} \end{cases} \quad (2.61)$$

The left-right consistency check is a suitable method to eliminate wrong disparity values which correspond to pixels which are occluded in one of the two cameras.

2.6.5 Disparity Map Filtering

After the initial disparity estimation, the quality of the disparity map can be improved by different post-processing steps. The cross-bilateral [Kopf07, Riemens09] is a suitable post-processing filter for the following tasks:

- align boundaries in the disparity with boundaries in the texture image;
- fill holes in the disparity map resulting from a left-right consistency check;
- upsample disparity maps from a lower resolution to the potentially higher resolution of the texture image.

Due to its wide spectrum of applications, the cross-bilateral filter will be described in more detail. The basis is a bilateral filter as proposed by [Tomasi98] which takes two terms into account for the pixel filtering: similarity in the color value and spatial proximity within a single gray scale or color image. Within the extension towards the cross-bilateral or joint-bilateral, the filter weights are calculated similarly to the simple bilateral filter but applied to another image, e.g. a depth image or disparity map. The filter weights $w(x_i, y_j)$ are calculated according to eqn. (2.62):

$$w(x_i, y_j) = \underbrace{e^{-\frac{(x_i - x_0)^2 + (y_j - y_0)^2}{2 \cdot \sigma_{\text{spatial}}^2}}}_{w_{\text{spatial}}(\Delta \text{Position})} \cdot \underbrace{e^{-\frac{(I(x_i, y_j) - I(x_0, y_0))^2}{2 \cdot \sigma_{\text{colour}}^2}}}_{w_{\text{color}}(\Delta \text{color})}. \quad (2.62)$$

The resulting weighting factor is the product of a spatial difference factor and a colorimetric difference factor.

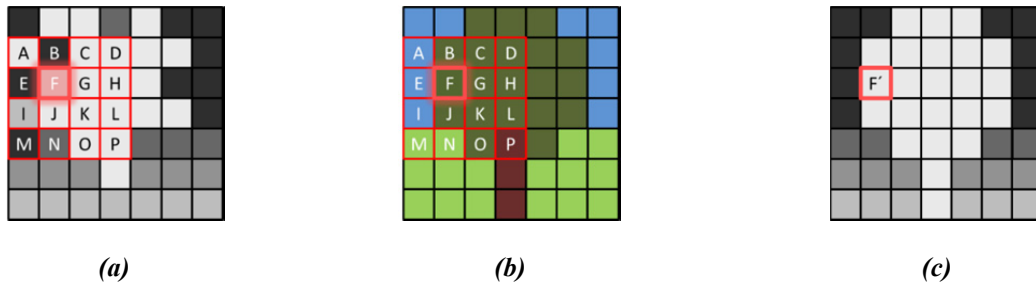


Figure 2.26. The edge preserving cross-bilateral filter can be used to align object boundaries with depth discontinuities, provided that the respective foreground and background objects have different colors. Moreover, this filtering concept can be used to up-sample the spatial resolution of disparity maps. Using the cross-bilateral filter, a high resolution colored texture image (b) can be combined with a noisy or subsampled disparity map (a) to generate a high resolution disparity map (c).

The concept is illustrated in the Figure 2.26. A noisy disparity map as shown in Figure 2.26 (a) which contains pixels with missing or wrong disparity value shall be aligned with a high resolution RGB color image from Figure 2.26 (b) to create a filtered disparity map as shown in Figure 2.26 (c). This process is performed pixel by pixel. In the example from Figure 2.26, the filtered target disparity F' shall be calculated for the pixel lying at the position F in the input image. In this context, the RGB values and corresponding disparities are evaluated in the neighborhood (i.e. filter window) of F. In this example, this filter window has a reduced size of 4x4 pixels to keep the figures clear and readable. In practice, the filter windows are much larger, while usually the pixel to be filtered is the center pixel of the filter window which therefore has typically an odd number of rows and columns.

The evaluation of the neighboring pixels consists of two steps. First, the color difference between the current pixel and each neighboring pixel is evaluated, usually in RGB space but it is also possible to perform this test in YUV or Lab color space. Now, the color difference is evaluated resulting in the filter weight w_{color} according to eqn. (2.62). If the colors are very similar, the corresponding weight is high. In the example from Figure 2.26 (b), the reference pixel F is dark green, hence as the pixels denoted as B, C, D, G, H etc. have the same color, the corresponding weights are high. They correspond to the same segment in the color image (the tree's leaves, from Figure 2.26 (b)) as the reference pixel F. The pixel denoted as P has a very different color (brown as the tree's trunk), hence the corresponding weight is low. The same holds true for the background pixels (blue sky, high difference, low weights) and with limits for the light green background (i.e. medium difference in color). A colorimetric separation of foreground and background is important for the functionality of the cross-bilateral filter as the color segmentation implicitly yields to a segmentation in disparity space which is a wanted result of the filtering operation.

During a second weighting operation the distance in pixels between the reference pixel F and the respective pixels within the filter kernel is evaluated leading to w_{spatial} which increases with spatial proximity. The direct neighbors B, E, J, K have a higher weight than the pixels D, L, M, O and P. Finally, according to eqn. (2.62), the two weights are multiplied to form the final weighting factor for the corresponding pixel. As a result, the cross-bilateral filter calculates the weighted average of the disparities.

2.7 Conclusion

In this chapter theoretical foundations which are required for the subsequent chapters were given. In a first step, concepts of 3D acquisition and reproduction were explained, such as positive and negative parallax, convergence plane and screen plane. Subsequently, an overview of important concepts regarding the relationship between the human visual system and depth perception were presented, e.g. binocular and monocular depth cues and possible 3D perception conflicts. The geometrical implications for stereoscopic 3D productions and important stereoscopic parameters such as the

adjustment of the inter-axial distance, the convergence distance and the mechanical alignment of the stereo rig were explained in a following step. The three sections 2.1 to 2.3 lay the fundament for the thesis in general and chapter 5 in particular.

In section 2.4, concepts of the projective geometry were derived, starting with basic concepts from single camera geometry such as the projection matrix. Subsequently, the concepts were first expanded towards stereo camera setups, introducing the fundamental matrix and giving an overview of stereo rectification algorithms. Finally, the concepts were extended towards three-camera geometry including foundations and related work of multi-camera rectification algorithms. The section 2.4 lays the fundament for chapter 3.

Basic concepts of feature detection and matching are described in section 2.5. The section starts with an overview of the process of interest point detection and related work. Concepts such as scale space pyramids, blob detectors and support regions were explained. Subsequently, different steps of the interest point description process along with an overview of related work was given. Finally, concepts relating to interest point matching were explained. The concepts described in section 2.5 lay the fundament for the proposed algorithm SKB presented in chapter 4.

In section 2.6, basic concepts of disparity estimation and filtering were explained. This included an overview of related work, a description of a typical stereo estimation workflow and related concepts such as similarity analysis, consistency checks and filtering e.g. using cross-bilateral filters. The concepts of section 2.6 will be used in chapter 6 where a multi-camera disparity estimation algorithm is described.

3 Linearized Projective Geometry

In this chapter, two main contributions of the thesis are presented, a new algorithm to estimate the fundamental matrix \mathbf{F} optimized for stereo camera setups and a new algorithm to estimate the trifocal tensor optimized for linear camera arrays¹³. In addition to the respective geometric entities, both algorithms deliver rectifying homographies for the respective camera setup. Both algorithms have in common that assumptions about the camera setup are made. This a priori knowledge makes the estimation process more robust. The assumption is that the setup is near the rectified state of the two or three-camera setup. This assumption is usually met when shooting a scene using a camera setup similar to the ones shown in Figure 1.2 and Figure 1.6. The mathematical approach for the estimation of the projective entities is to develop a Taylor expansion around the ideal state which will be truncated after the linear term.

Related work regarding stereo image rectification and estimation of the fundamental matrix has been presented in section 2.4.2.2, while the related work concerning trifocal image rectification was presented in section 2.4.3.1.

3.1 Taylor Expansion for Projective Entities

The concept of the linearization by means of a truncated Taylor expansion can be applied to large set of geometric entities and can be thought as a special mathematical approach towards projective geometry. The concept for linearization with respect to projective entities followed in this thesis is based on the approach of Shi and Tomasi who applied the concept in [Shi94] to calculate a homography representing an affine motion. Linearization is in fact a common and valuable approach not only within the projective geometry, but in nearly all fields of engineering. The motivation to linearize a mathematical problem can be pure necessity as no closed form solution might exist, or the underlying problem is far too complex in the general form. On the other hand, researchers might choose this path even in presence of an acceptable mathematical solution to take advantage of other aspects. In fact, the process of linearization is characterized by establishing a geometric model of the problem which is ideal and easy to describe and to assume that deviations between this model and the real data are small. This allows one to take most advantage of a priori knowledge. Thereby, the mathematical approach can be customized to an individual problem statement. This makes also the computation simpler and more robust in the case of noisy input data as it is possible to reduce the number of unknown variables (e.g. the focal length) if a priori information can be used. The basic concept of the linearization process will now be presented.

¹³ Parts of the content in this chapter have been previously published in [Zilly10a], [Zilly12c] and [Zilly14].

The Taylor expansion of a function $f(x, y)$ around the point $(0,0)$ is given by

$$f(x, y) = f(0,0) + x \frac{\partial f}{\partial x}(0,0) + y \frac{\partial f}{\partial y}(0,0) + \dots \quad (3.1)$$

where all terms of second or higher order have been omitted. The linearized result is the sum of the function at a known point (e.g. the rectified state) and the partial derivatives at this point multiplied by the deviation from the known state. In a similar way, it is possible to express a rotation matrix as a function of three variables, e.g. the angles α_x , α_y , and α_z . To clarify the linearization approach, the calculation of the first order Taylor expansion of the rotation matrices is performed explicitly. The rotation matrix $R(\alpha_x, \alpha_y, \alpha_z)$ from eqn. (2.18) is a 3×3 matrix which is composed of trigonometric functions. However, it can be assumed that all angles α_x , α_y , and α_z are small near the rectified state:

$$\begin{aligned} R(\alpha_x, \alpha_y, \alpha_z) &\approx \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \alpha_x \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} + \alpha_y \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} + \alpha_z \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -\alpha_z & \alpha_y \\ \alpha_z & 1 & -\alpha_x \\ -\alpha_y & \alpha_x & 1 \end{bmatrix}. \end{aligned} \quad (3.2)$$

In the following, the concept will be extended to projection matrices, the fundamental matrix and finally the trifocal tensor.

3.2 Linearized Computation of the Fundamental matrix

An optimal stereo sequence needs to be rectified in order to avoid vertical disparities and similar image distortions. This requirement is in general not met without electronic post-processing due to the finite accuracy of mechanical components. On the other hand, it can be expected that the two cameras are near the rectified state, because otherwise, the content would be unsuitable to be watched in 3D.

The relation between the fundamental matrix and stereo rectification along with different existing rectification methods and related work has been mentioned in section 2.4.2. The method proposed in this section was first published in [Zilly10a]. It serves as related work for a rectification method proposed by Georgiev et al. in [Georgiev13]. The authors from [Georgiev13] also compare their method to the one from [Zilly10a]. Compared to the methods presented in section 2.4.2, the method presented in this section has a better alignment performance as allows for a simple pose estimation of the stereo cameras.

Most rectification methods are based on a strong calibration where intrinsic and extrinsic parameters are known. However, calibration data is often not provided in the context of a 3D shooting, such that the rectification needs to be done using point correspondences. In this section, a rectification technique

which estimates the fundamental matrix jointly with the appropriate rectification parameters is proposed. The algorithm is designed for narrow baseline stereo rigs. It is assumed that the optical axes are almost parallel beside a possible convergence angle. The rectification parameters allow a pose estimation of one camera relative to the other one so that the mechanical alignment of the stereo rig can be improved. An overview of related stereo rectification algorithms was given in section 2.4.2.2.

The method proposed in this thesis establishes a relationship between the components of the fundamental matrix, and a physical model of the camera positions. This allows calculating the rectifying homographies with a very small distortion impact. The model assumes a geometry which is near the rectified state such that the fundamental matrix can be expressed as linearized Taylor expansion around the rectified state as defined in eqn. (2.31).

3.2.1 Point Correspondences

To establish the point correspondences, which are used for the estimation process, a robust feature detector is needed which produces as few outliers as possible. Suitable feature detectors are SIFT [Lowe04] combined with Difference of Gaussian interest point detection and Up-Right-SURF [Bay08] and the Hessian Box-Filter detector. The basic principles of feature detection and matching along with an overview of state-of-the-art feature detectors were given in 2.5. Moreover, a new feature descriptor called SKB will be presented in chapter 4. However, even very distinctive descriptors will produce a certain amount of outliers. One well known technique is to eliminate outliers using a RANSAC estimation [Fischler80] of the fundamental matrix [Hartley06]. For the remainder of this chapter, it is assumed that a list of point correspondences $\mathbf{m} = (u, v, 1)^T$ and $\mathbf{m}' = (u', v', 1)^T$ as introduced in section 2.4.2 is available.

3.2.2 Linearization Approach

The aim is to develop a Taylor expansion of the fundamental matrix around the rectified state. In order to linearize the algorithm, the Taylor expansion is cut after the first term. The fundamental matrix can be expressed using the terms \mathbf{R}' , \mathbf{K}' , \mathbf{K} , and \mathbf{t} from eqn. (2.26). The latter component \mathbf{t} is the translation vector as defined in eqn. (2.28). The nomenclature for the cross-product representation is defined in eqn. (2.27) while basic principles of the epipolar geometry have been introduced in section 2.4.2.

All components of the fundamental matrix will be linearized separately, and after multiplying them according to eqn. (2.26), all second order terms will be omitted. This yields to the same result which one would get when multiplying all non-linearized terms and calculating all mixed derivatives afterwards. The stratified calculation is chosen here because it is better suited to understand the concept of the approach.

The fundamental matrix \mathbf{F} is defined only up to a scale. Hence, eqn. (2.26) is still valid if the translation vector $\mathbf{t} = (t_x, t_y, t_z)^T$ is divided by t_x . The result is denoted $\hat{\mathbf{t}}$ with $\hat{\mathbf{t}} = (1, \hat{t}_y, \hat{t}_z)^T$. It can

be assumed that the camera setup is near the rectified state, hence the translation in y-direction or z-direction is small compared to the horizontal displacement of the two cameras. It can be concluded that $\hat{t}_y \ll 1$ and $\hat{t}_z \ll 1$. The resulting translation vector in cross product representation is given by the following matrix:

$$[\hat{\mathbf{t}}]_{\times} = \begin{bmatrix} 0 & -\hat{t}_z & \hat{t}_y \\ \hat{t}_z & 0 & -1 \\ -\hat{t}_y & 1 & 0 \end{bmatrix}. \quad (3.3)$$

Assuming that the rotation angles are small, all second order derivatives are neglected ($\alpha \ll 1$) according to (3.1) which yields to the following rotation matrix similar to the one in eqn. (3.2):

$$\widehat{\mathbf{R}}' = \begin{bmatrix} 1 & -\alpha_z & \alpha_y \\ \alpha_z & 1 & -\alpha_x \\ -\alpha_y & \alpha_x & 1 \end{bmatrix}. \quad (3.4)$$

By multiplying $[\hat{\mathbf{t}}]_{\times}$ by $\widehat{\mathbf{R}}'$ and eliminating any mixed term as second order effect, the following linearized essential matrix \mathbf{E} results:

$$\mathbf{E} = \begin{bmatrix} 0 & -\hat{t}_z & \hat{t}_y \\ \hat{t}_z + \alpha_y & -\alpha_x & -1 \\ -\hat{t}_y + \alpha_z & 1 & -\alpha_x \end{bmatrix}. \quad (3.5)$$

Concerning the intrinsic matrices it is assumed that the principal point is centered, that the aspect ratio is 1 and that the skew is zero while the focal lengths f and f' might differ. The ratio of the focal lengths will be denoted as $f'/f = 1 + \alpha_f$ where $\alpha_f \ll 1$ assuming that the deviation between the two focal lengths is small compared to f . This leads to the following matrices \mathbf{K}^{-1} and \mathbf{K}'^{-1} respectively for the left and right camera:

$$\mathbf{K}^{-1} = \begin{bmatrix} 1/f & 0 & 0 \\ 0 & 1/f & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (3.6)$$

$$\mathbf{K}'^{-1} = \begin{bmatrix} \frac{1-\alpha_f}{f} & 0 & 0 \\ 0 & \frac{1-\alpha_f}{f} & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (3.7)$$

Please note that the origin lies in the image center and that the term $1/(1 + \alpha_f)$ can be approximated by with $1 - \alpha_f$ as long as α_f is small. Inserting the results from eqns. (3.3)-(3.7) into eqn. (2.26) leads to the following linearized fundamental matrix \mathbf{F} :

$$\mathbf{F} = \begin{bmatrix} 0 & -\frac{\hat{t}_z}{f} & \hat{t}_y \\ \frac{\hat{t}_z + \alpha_y}{f} & -\frac{\alpha_x}{f} & -1 + \alpha_f \\ -\hat{t}_y + \alpha_z & 1 & -f \cdot \alpha_x \end{bmatrix} \quad (3.8)$$

Second order effects (e.g. $\alpha_f \hat{t}_y = 0, \alpha_f \alpha_x = 0, \dots$) have been eliminated while the result has been multiplied by f . The translation vector \mathbf{t} is substituted according to eqn. (2.28): $\hat{t}_y = \hat{c}_y + \alpha_z$ and $\hat{t}_z = -\hat{c}_z + \alpha_y$. The result is the Taylor expansion of fundamental matrix which has been truncated after the linear term as shown in eqn. (3.9). The linearized fundamental matrix from eqn. (3.9) differs from the rectified fundamental matrix from eqn. (2.31) only by linear terms.

$$\mathbf{F} = \begin{bmatrix} 0 & \frac{-\hat{c}_z + \alpha_y}{f} & \hat{c}_y + \alpha_z \\ \frac{\hat{c}_z}{f} & -\frac{\alpha_x}{f} & -1 + \alpha_f \\ -\hat{c}_y & 1 & -f \cdot \alpha_x \end{bmatrix} \quad (3.9)$$

In the next sub-section, the estimation process for the linearized fundamental matrix will be described.

3.2.3 Estimation of the Linearized Fundamental Matrix

In the previous sub-section, a formula for the linearized fundamental matrix was derived. During the following estimation process, the geometric components such as α_f or \hat{c}_y will be calculated individually. The first step is to insert the linearized fundamental matrix \mathbf{F} from eqn. (3.9) into the epipolar constraint from eqn. (2.25). It is assumed that the required point correspondences $\mathbf{m} = (u, v, 1)^T$ and $\mathbf{m}' = (u', v', 1)^T$ were provided. The epipolar constraint transforms to the following equation:

$$u' \left(v \frac{-\hat{c}_z + \alpha_y}{f} + \hat{c}_y + \alpha_z \right) + v' \left(u \frac{\hat{c}_z}{f} - v \frac{\alpha_x}{f} - 1 + \alpha_f \right) + (u(-\hat{c}_y) + v - f \cdot \alpha_x) = 0. \quad (3.10)$$

The above equation can be regrouped such that the vertical disparity $v' - v$ is on the left and all terms inducing it on the right. After regrouping the equation by terms inducing vertical disparities the following simple equation results:

$$\underbrace{v' - v}_{\text{vert. disp.}} = \underbrace{\hat{c}_y(u' - u)}_{y\text{-error}} + \underbrace{\alpha_z u'}_{\text{roll}} + \underbrace{\alpha_f v'}_{\text{zoom}} + \underbrace{\alpha_x}_{\text{mism.}} + \underbrace{-f \cdot \alpha_x}_{\text{tilt}} + \underbrace{\alpha_y u' v / f}_{\text{keystone}} + \underbrace{-\alpha_x v v' / f}_{\text{tilt keyst.}} + \underbrace{\hat{c}_z (u v' - u' v) / f}_{z\text{-error}}. \quad (3.11)$$

It is now possible to build up a system of linear equations which enables calculating the coefficients which are needed to compose the fundamental matrix. A constraint matrix \mathbf{A} is built using point correspondences between the left and the right camera. Assuming that $\mathbf{m} = (u, v, 1)^T$ and $\mathbf{m}' =$

$(u', v', 1)^T$ denote the i^{th} point correspondence, then the i^{th} row of the constraint matrix has the following form:

$$\mathbf{A}_i = (1, u', v', u' - u, u'v, vv', uv' - u'v) \quad (3.12)$$

while the i^{th} component of the constraint vector \mathbf{b} is:

$$\mathbf{b}_i = \mathbf{v}' - \mathbf{v}. \quad (3.13)$$

The system of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ can now be solved using the pseudo-inverse of \mathbf{A} :

$$\mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}. \quad (3.14)$$

The result vector \mathbf{x} contains the coefficients which can now be used to compose the fundamental matrix from eqn. (3.9) and to calculate rectifying homographies:

$$\mathbf{x} = (-f \cdot \alpha_x, \alpha_z, \alpha_f, \hat{c}_y, \alpha_y/f, -\alpha_x/f, \hat{c}_z/f)^T. \quad (3.15)$$

3.2.4 Choice of the Fitting Parameters

A constrained fundamental matrix has been estimated which is composed of a set of meaningful parameters, i.e. angles, offsets in pixel, difference in focal length. It is possible to omit one or more parameters to reduce the number of unknown variables. For instance, the estimation of \hat{c}_z depends on four coordinates which makes this estimation numerically unstable. Furthermore, f can be deduced by the two tilt coefficients, however, when the tilt angle α_x vanishes, the estimation of f is numerically unstable. It might be a good choice to omit the estimation of \hat{c}_z and possibly the estimation of $-\alpha_x/f$. The latter parameter might be neglected when the vertical opening angle is small, e.g. when tele-zoom lenses are used. Furthermore, any pre-knowledge of the geometric setup can be exploited. If, for instance, it is known that the focal lengths match perfectly, i.e. $\alpha_f = 0$, this coefficient can be omitted as well. With the same argument, one might omit the estimation of the toe-in α_y if, for instance, the image pair is already de-keystoned. This linearized approach allows for a fine granular control of the estimation performance and gives also an insight of the sources of numerical unstableness.

3.2.5 Model Fitting with RANSAC

The estimation process described in section 3.2.3 can be improved using a RANSAC fitting [Fischler80] similar to the estimation of the fundamental matrix described in [Hartley04]. The approach tries to eliminate inconsistent feature point pairs. In this context, a set of point

correspondences denoted as *sample* is randomly chosen from the complete list of feature points. Now, a temporary fundamental matrix \mathbf{F}_{temp} is estimated using this sample. Subsequently, the Sampson distance defined in eqn. (2.29) is calculated for the complete list of point correspondences using \mathbf{F}_{temp} . Finally, the number of inliers for which the Sampson distance is below a threshold $\theta_{sampson}$ is counted. The process will be repeated for a predefined number of iterations or until the percentage of inliers exceeds a predefined threshold while choosing a new *sample* for each iteration. After the completion of all iterations, the temporary fundamental matrix which generated the highest number of inliers is retrieved as best fundamental matrix \mathbf{F}_{best} . Finally, those feature point pairs which are consistent with \mathbf{F}_{best} are identified and this list is used to build the constraint matrix \mathbf{A} which yields to the final fundamental matrix \mathbf{F}_{final} . The matrix \mathbf{F}_{final} is estimated using a high number of feature point pairs and is therefore less sensitive to noise than the matrix \mathbf{F}_{best} which was estimated only using the feature point pairs within the respective *sample*.

An important question is the ratio of inliers compared to outliers within the feature point pairs. Obviously, at least a single sample should be free of outliers, otherwise the RANSAC process yields to unsatisfactory results [Hartley04]. Please note that the feature point descriptor SKB presented in chapter 4 is designed to achieve a low outlier rate.

The sample size s plays an important role in the number of needed samples for the RANSAC, especially when the percentage of outliers is high. The minimum number of samples or iterations required to have with a probability of p at least one sample with inliers only given an outlier rate ϵ is:

$$N = \left\lceil \frac{\log(1 - p)}{\log(1 - (1 - \epsilon)^s)} \right\rceil. \quad (3.16)$$

The following table illustrates this [Hartley04] for different sample sizes and an assumed proportion of outliers $\epsilon = 50\%$ and $p = 99.9\%$. It is assumed that such a high probability p is needed to guarantee a satisfactory rectification.

Table 3.1. Minimum number of required samples for a RANSAC fit of the linearized fundamental matrix

Sample size s	3	4	5	6	7
Required Samples	52	108	218	439	881

If for instance two of the seven parameters were omitted, the sample size s reduces to 5. Consequently, the number of required point correspondences required to estimate the linearized fundamental matrix reduces from 881 to 218, assuming an outlier rate $\epsilon = 50\%$. It can be deduced that the possibility to adapt the estimation process by using a priori information greatly improves the robustness of the overall algorithms.

3.2.6 Singularity Constraint

The fundamental matrix has rank 2 and hence the determinant should be zero. If the assumption of vanishing second order effects is correct (which of course is the case only up to a certain precision), then the equation (3.9) should lead to a vanishing determinant. However, the numerical value will be non-zero and can be interpreted as an indicator, how well the model of the linearization works. The singularity constraint will not be enforced using the SVD method as described in [Hartley04]. In fact, the direct relationship between the components of the fundamental matrix and their physical interpretation as described by (3.9) would be lost when enforcing the singularity constraint.

3.2.7 Rectifying Homographies

Once the geometric parameters from eqn. (3.15) have been calculated, the rectifying homographies can directly be computed. The angles associated with roll and tilt and convergence, i.e. $\alpha_z, \alpha_x, \alpha_y$, can be corrected by rotating \mathbf{P}' in the inverse direction. A deviation of \hat{c}_y can be corrected by rotating both cameras around the z-axis in the same direction by an angle given by $\text{atan}(\hat{c}_y/\hat{c}_x)$. Taking into account that \hat{c}_y is small compared to \hat{c}_x which itself has been normalized to 1, the resulting angle can be approximated by \hat{c}_y . A deviation of \hat{c}_z can be corrected by rotating both cameras around the y-axis in the same direction by an amount \hat{c}_z . The difference of the two focal lengths denoted as $\Delta - \text{Zoom}$ or zoom mismatch can be corrected using the following homography:

$$\mathbf{H}'_{\Delta-\text{Zoom}} = \begin{bmatrix} 1 - \alpha_f & 0 & 0 \\ 0 & 1 - \alpha_f & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (3.17)$$

The rectifying homographies have the form:

$$\mathbf{H} = \mathbf{K} \cdot \mathbf{R}^T \cdot \mathbf{K}^{-1} \quad (3.18)$$

for the left camera and

$$\mathbf{H}' = \mathbf{K} \cdot \mathbf{R}^T \cdot \mathbf{K}'^{-1} \quad (3.19)$$

for the right camera. This ensures normalized intrinsic matrices and the correct orientation of both cameras after applying the homographies. The following homographies are a linearized approximation of the ideal homographies. The advantage is that they can be composed directly from the result vector \mathbf{x} from eqn. (3.15). The rectifying homography for the left view is:

$$\mathbf{H} = \begin{bmatrix} 1 & \hat{c}_y & 0 \\ -\hat{c}_y & 1 & 0 \\ -\hat{c}_z/f & 0 & 1 \end{bmatrix}, \quad (3.20)$$

while the homography for the right view is:

$$\mathbf{H}' = \begin{bmatrix} 1 - \alpha_f & \alpha_z + \hat{c}_y & 0 \\ -(\alpha_z + \hat{c}_y) & 1 - \alpha_f & -f\alpha_x \\ (\alpha_y - \hat{c}_z)/f & -\alpha_x/f & 1 \end{bmatrix}. \quad (3.21)$$

Disparities for objects in the plane which induces the homographies vanish [Fusiello08]. The rectification can be done with respect to the plane at infinity, yielding to parallel optical axes. In that case, the upper-right entry of \mathbf{H} should be $f\hat{c}_z$ and the upper right entry of \mathbf{H}' should be $f(\alpha_y - \hat{c}_z)$. The upper-right entries of the rectifying homographies induce an offset of the horizontal disparities which can be verified by simple calculus. However, parallel optical axes would lead to a convergence plane which is shifted to infinity. As the convergence is an important creative parameter during stereoscopic 3D production, a rectification which is optimized for this application should preserve the convergence plane. Consequently, the upper-right entries of \mathbf{H} and \mathbf{H}' are set to zero which means that following the notation used in [Fusiello08] the rectification is done with respect to the current convergence plane. Different concepts of 3D reproduction explaining the role of the convergence plane are described in section 2.1 through section 2.3.

3.2.8 Results

3.2.8.1 Comparison to Mallon and Georgiev

In a first experiment, the method was applied to the dataset supplied by Mallon and Whelan [Mallon05]¹⁴. Six image pairs were rectified using the point correspondences provided within the dataset. The point correspondences were inserted into the system of linear equations according to (3.12). To ensure that the results can be compared with [Mallon05] and [Georgiev13], all point correspondences were used, without a prior RANSAC filtering. Subsequently, the result vector \mathbf{x} from eqn. (3.15) which contains the coefficients describing the epipolar geometry was computed. Using the nomenclature from eqn. (3.11), the best performance was achieved when fitting for the following six coefficients: *y - error*, *roll*, *zoom mismatch*, *tilt*, *keystone*, and *tilt keyst*. These coefficients were subsequently used to build the homographies \mathbf{H} and \mathbf{H}' according to (3.20) and (3.21). The resulting homographies were then used to rectify the image pairs. The original and the rectified image pairs are shown in Figure 3.1. In order to perform a quantitative comparison of the rectification results, a set of error metric parameters were computed following [Mallon05]. For each homography, measures for orthogonality E_0 (ideally 90°), aspect ratio E_a (ideally 1.0), and rectification error E_r were computed. The values obtained with the proposed method are shown together with the data provided by [Mallon05] in Table 3.2. The results regarding Mallon's, Loop's, and Hartley's method were transferred from [Mallon05] and the result regarding Georgiev from [Georgiev13].

¹⁴ Available from <http://elm.eeng.dcu.ie/vsl/vsgcode.html>

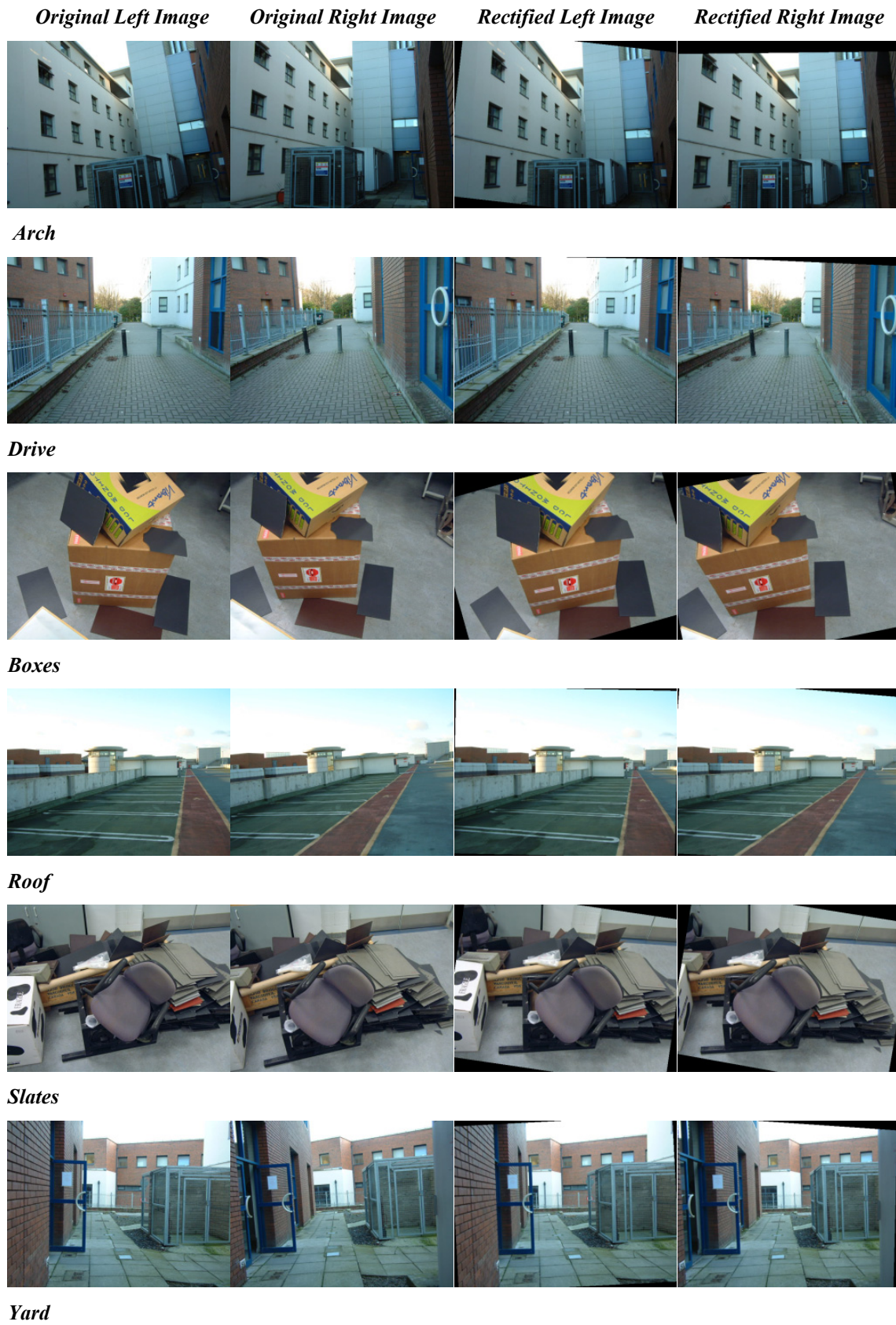


Figure 3.1. Visualization of the proposed rectification method. Original images retrieved from [Mallon05].

Table 3.2. Comparison of different rectification techniques. The results regarding Mallon, Loop, and Hartley were transferred from [Mallon05], the results for Georgiev from [Georgiev13] with fewer digits for the column E_a . The sample Boxes was not investigated in [Georgiev13].

Sample	Method	E_0		E_a		Error E_r	
		H'	H	H'	H	Mean	Std
Arch	Proposed	89.96	90.00	0.9988	1.0000	0.14	0.36
	Georgiev	89.71	89.92	0.99	1.00	0.23	0.23
	Mallon	91.22	90.26	1.0175	1.0045	0.22	0.33
	Loop	95.40	98.94	1.0991	1.1662	131.3	20.63
	Hartley	100.74	93.05	1.2077	1.0546	39.21	13.85
Drive	Proposed	89.98	90.00	0.9992	1.0000	0.01	0.93
	Georgiev	89.92	90.03	1.00	1.00	0.37	0.71
	Mallon	90.44	90.12	1.0060	1.0021	0.18	0.91
	Loop	98.73	101.42	1.1541	1.2052	10.41	3.24
	Hartley	107.66	90.87	1.3491	1.015	3.57	3.43
Boxes	Proposed	90.02	90.00	1.0000	1.0000	0.18	0.52
	Georgiev	n/a	n/a	n/a	n/a	n/a	n/a
	Mallon	88.78	89.33	0.9785	0.9889	0.44	0.33
	Loop	97.77	95.69	1.1279	1.0900	4.35	9.20
	Hartley	86.56	94.99	0.9412	1.0846	33.36	8.65
Roof	Proposed	90.01	90.00	1.0009	1.0000	0.06	1.15
	Georgiev	89.98	90.09	1.00	1.00	0.44	0.10
	Mallon	88.35	88.23	1.1077	0.9700	1.96	2.95
	Loop	69.28	87.70	0.6665	1.0497	0.84	11.01
	Hartley	122.77	80.89	1.5256	0.8552	11.89	18.15
Slates	Proposed	90.00	90.00	1.0001	1.0000	0.23	0.20
	Georgiev	90.43	90.58	1.01	1.02	0.13	0.15
	Mallon	89.12	89.13	0.9852	0.9855	0.59	0.56
	Loop	37.29	37.15	0.2698	0.2805	1.14	3.84
	Hartley	89.96	88.54	1.0000	0.9769	2.27	5.18
Yard	Proposed	90.05	90.00	1.0024	1.0000	0.12	0.44
	Georgiev	90.01	89.89	1.00	1.00	0.31	0.28
	Mallon	89.91	90.26	0.9987	1.0045	0.53	0.54
	Loop	133.62	134.27	2.1477	2.4045	8.91	13.19
	Hartley	101.95	91.91	1.2303	1.0335	48.19	11.49

Table 3.2 shows that for the proposed method, the image distortions measured by E_0 and E_a are considerably smaller than for any other method. The homography H has always orthogonality $E_0 = 90$ and aspect ratio $E_a = 1.0$. No shearing or anisotropic scaling was induced by H , only a rotation around the image center. The values for H' indicate a very low image distortion. Concerning the rectification error E_r , the proposed method shows good alignment performance. The mean of E_r is nearer to 0 for every image pair. The standard deviation shows an accuracy similar to Mallon's and Georgiev's method.

3.2.8.2 Rectification including F-matrix estimation

In a second experiment, a frame from the Beergarden [3D4YOU] stereo sequence was selected. The SIFT feature detector was used to find putative matches [Lowe04]. Subsequently, the RANSAC filtering described in 3.2.5 was used to eliminate outliers. To perform one RANSAC iteration, four point correspondences were used to fill the constraints matrix \mathbf{A} using eqn. (3.12) and to fit the results vector \mathbf{x} from eqn. (3.15) including y – error, roll, zoom mismatch and tilt. Afterwards, these values were used to compute the candidate fundamental matrix \mathbf{F} according to eqn. (3.9).

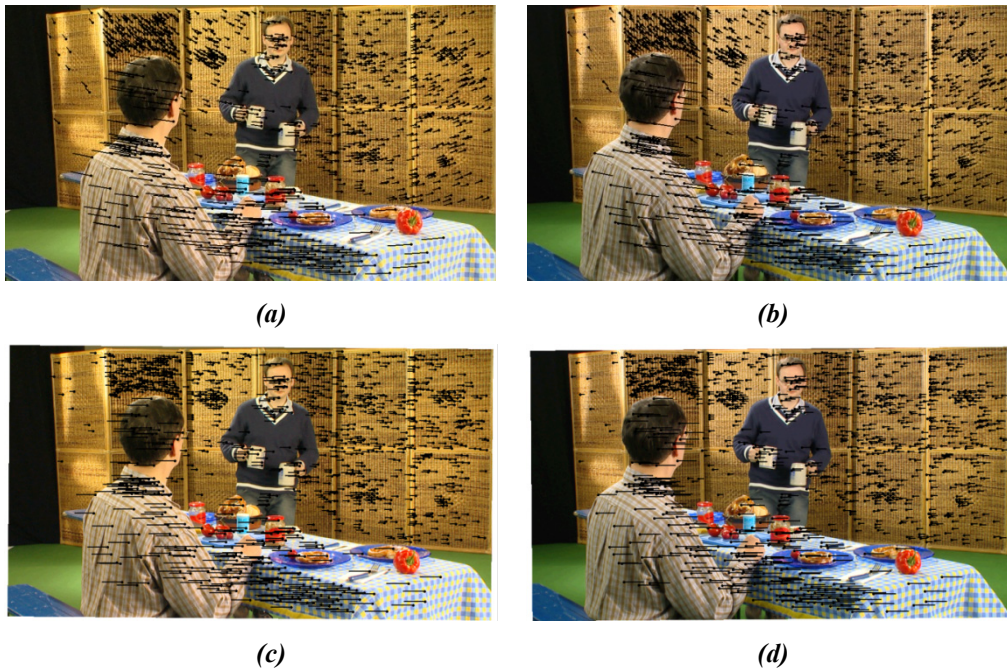


Figure 3.2. original left (a), original right (b), rectified left (c), rectified right (d). The black arrows indicate the horizontal and vertical disparities of feature points matched between the left and right images. Vertical disparities have been eliminated after the rectification process.

Figure 3.2 shows the inlying matches for the left and the right image before (a,b) and after (c,d) rectification. In Figure 3.3 the original images and the rectified images are overlaid which allows for a qualitative evaluation of the rectification performance.



Figure 3.3. A stereo pair from the Beergarden Sequence shot in the 3D4YOU Project [3D4YOU]. Overlay of the two original images (a) and the rectified images (b).

The room divider in the background allows for a good inspection of the alignment of the two cameras. Apparently, the rectification process resulted in a well aligned image pair. The proposed rectification algorithm will be used inside the assistance system described in chapter 5.

3.3 Linearized Computation of the Trifocal Tensor

In the previous sub-section, an approach for the linearized computation of the fundamental matrix has been proposed. In this sub-section, the concept will be extended towards the linearized computation of the trifocal tensor. In a first step, a geometric setup is defined which can be regarded as the ideal setup. The trifocal tensor associated to this setup can easily be derived. Subsequently, a Taylor expansion of this function around this setup can be performed which is cut after the linear term. More precisely, the trifocal tensor can then be interpreted as a multi-variate function while the ideal geometric setup is defined by a set of coordinates consisting of all required variables.

As the trifocal tensor represents the geometry between three cameras it can be used to calibrate a multi-camera system, and to derive rectifying homographies. The algorithm achieves vertical alignment and horizontal alignment. In contrast to the related work presented in section 2.4.3.1 the method presented in this chapter is able to ensure a horizontal alignment as well as a vertical alignment. After rectification, horizontal disparities are proportional to each other up to a proportionality constant β , i.e. the ratio of the camera baselines. The main properties of the proposed rectification method are the following:

- Based on feature point triplets,
- Suitable for uncalibrated cameras in a linear but possibly non-equidistant configuration,
- Elimination of vertical disparities,
- Horizontal disparities become proportional to the ratio β of the two camera baselines,
- Provision of the proportionality constant β ,
- Linear estimation of the trifocal tensor,
- Robust against noise and outliers.

In the following, the assumptions regarding the geometric setup and details of the proposed algorithm for the estimation of the trifocal tensor are presented.

3.3.1 Ideal Geometric Setup

Different geometric setups are possible with three cameras. Beside the case where all three cameras are in general position, two setups are of special interest for computer vision applications: the L-shaped approach as mentioned in 2.4.3.1 and the linear camera array where all camera centers are on a common baseline. For the remainder of this chapter, the latter approach will be investigated.

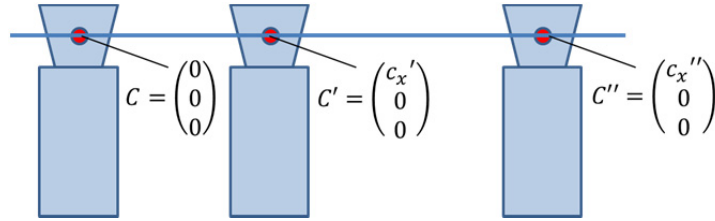


Figure 3.4. Trifocal setup around which the Taylor expansion will be developed.

Figure 3.4 shows the trifocal setup around which the Taylor expansion will be developed. All cameras' centers C , C' and C'' lie on a common baseline while all optical axes are parallel. Please note that this is not necessarily an equidistant setup. In the ideal geometric setup, all cameras are normalized. The resulting projection matrices P , P' , and P'' are as follows:

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, P' = \begin{bmatrix} 1 & 0 & 0 & -c'_x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, P'' = \begin{bmatrix} 1 & 0 & 0 & -c''_x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \quad (3.22)$$

Using eqns. (2.38) and (2.39) it is possible to calculate the trifocal tensor associated to the ideal state around which the Taylor expansion will be developed:

$$T_1 = \begin{bmatrix} c'_x - c''_x & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad T_2 = \begin{bmatrix} 0 & c'_x & 0 \\ -c''_x & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad T_3 = \begin{bmatrix} 0 & 0 & c'_x \\ 0 & 0 & 0 \\ -c''_x & 0 & 0 \end{bmatrix}. \quad (3.23)$$

In this ideal case, disparities calculated from corresponding point triplets $\{m, m', m''\}$ follow a simple rule: corresponding pixels lie in the same image scanline in all three views. Horizontal disparities are proportional to each other, i.e. for all corresponding point triplets with $\mathbf{m} = (u, v, 1)$, $\mathbf{m}' = (u', v', 1)$ and $\mathbf{m}'' = (u'', v'', 1)$ visible in the three cameras, the horizontal coordinates obey the simple equation:

$$u'' - u = \beta \cdot (u' - u) \quad (3.24)$$

where the proportionality constant β is the ratio of the camera baselines.

$$\beta = c_x''/c_x'. \quad (3.25)$$

Also in the ideal case, the vertical disparities are 0, along the three views, i.e.

$$v'' - v = v' - v = 0. \quad (3.26)$$

In the general case, a trifocal consistency check is much more complex and inhibits the use of the trifocal tensor \mathcal{T} to perform a point transfer.

3.3.2 Degrees of Freedom for the Linearized Geometric Setup

In the following, the degrees of freedom of the linearized geometric setup will be introduced, i.e. by what deviations from the ideal state the model will be enhanced. In a first step, enhanced intrinsic matrices \mathbf{K} , \mathbf{K}' , and \mathbf{K}'' will be developed along with enhanced rotation matrices \mathbf{R} , \mathbf{R}' , and \mathbf{R}'' . These will be used to calculate the projection matrices which will finally be used to calculate the trifocal tensor.

Without loss of generality, the center of origin will be placed in the reference camera, which the leftmost camera in Figure 3.4. The center of this camera can be denoted as $\mathbf{C} = (0,0,0)$. The remaining camera centers lie on a common baseline, its centers are $\mathbf{C}' = (c_x', 0, 0)$ and $\mathbf{C}'' = (c_x'', 0, 0)$ respectively, i.e. no translational errors are taken into account. In contrast, at least small deviations of the camera orientations of the second and third camera of the triplet as well as small deviations of the intrinsic parameters have to be taken into account. The former can be expressed using rotation matrices with three angles per camera, i.e. $\mathbf{R}'(\alpha_x', \alpha_y', \alpha_z')$ for the second and $\mathbf{R}''(\alpha_x'', \alpha_y'', \alpha_z'')$ for the third camera of the triplet according to eqn. (3.2). Please note that the linearization approach is similar to the two-camera case explained in section 3.2.2. The intrinsic parameters will be parameterized as follows: the camera matrix for the reference camera is the identity matrix, i.e. $\mathbf{K} = \mathbf{I}$. It is assumed that the focal lengths of the second and third camera differ by a small amount from the (normalized) focal length of the reference camera which will be denoted as α_f' for the second and α_f'' for the third camera. Moreover, a parameter α_r'' will be introduced which accounts for a small deviation of the aspect ratio from 1.0 of the third camera. Finally, with p_x'' an offset of the principal of the third camera from the image center will be parameterized. The latter two parameters are needed to ensure that after a multi-camera rectification, the horizontal disparities are proportional according to eqn. (3.24). The updated intrinsic matrices \mathbf{K}' and \mathbf{K}'' are defined as follows:

$$\mathbf{K}'(\alpha_f') = \begin{bmatrix} (1 + \alpha_f') & 0 & 0 \\ 0 & (1 + \alpha_f') & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{K}''(\alpha_f'', \alpha_r'', p_x'') = \begin{bmatrix} (1 + \alpha_f'')(1 + \alpha_r'') & 0 & p_x'' \\ 0 & (1 + \alpha_f'') & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (3.27)$$

This result will now be used to calculate the linearized projection matrices $\widehat{\mathbf{P}}'$ and $\widehat{\mathbf{P}}''$. In a first step, the projection matrices \mathbf{P}' and \mathbf{P}'' will be calculated using eqn. (2.21):

$$\mathbf{P}' = \mathbf{K}'(\alpha'_f) \cdot \mathbf{R}'(\alpha'_x, \alpha'_y, \alpha'_z) \cdot \begin{bmatrix} 1 & 0 & 0 & -c'_x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad (3.28)$$

$$\mathbf{P}'' = \mathbf{K}''(\alpha''_f, \alpha''_r, p''_x) \cdot \mathbf{R}''(\alpha''_x, \alpha''_y, \alpha''_z) \cdot \begin{bmatrix} 1 & 0 & 0 & -c''_x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Subsequently, all terms will be multiplied and second order terms will be eliminated according to the concepts explained in sub-section 3.1. This results in two linearized projection matrices:

$$\begin{aligned} \widehat{\mathbf{P}}' &= \begin{bmatrix} \alpha'_f + 1 & -\alpha'_z & \alpha'_y & -c'_x(\alpha'_f + 1) \\ \alpha'_z & \alpha'_f + 1 & -\alpha'_x & -c'_x\alpha'_z \\ -\alpha'_y & \alpha'_x & 1 & c'_x\alpha'_y \end{bmatrix}, \\ \widehat{\mathbf{P}}'' &= \begin{bmatrix} \alpha''_f + \alpha''_r + 1 & -\alpha''_z & \alpha''_y + p''_x & -c''_x(\alpha''_f + \alpha''_r + 1) \\ \alpha''_z & \alpha''_f + 1 & -\alpha''_x & -c''_x\alpha''_z \\ -\alpha''_y & \alpha''_x & 1 & c''_x\alpha''_y \end{bmatrix}. \end{aligned} \quad (3.29)$$

Finally the above linearization concept is applied to derive the trifocal tensor using \mathbf{P} , $\widehat{\mathbf{P}}'$, $\widehat{\mathbf{P}}''$ and eqns. (2.38) and (2.39). The result is the linearized trifocal tensor in slices representation:

$$\begin{aligned} \mathcal{T}_1 &= \begin{bmatrix} (1 + \alpha'_f + \alpha''_f + \alpha''_r)\Delta c_x & \alpha''_z\Delta c_x & -\alpha''_y\Delta c_x \\ \alpha'_z\Delta c_x + c'_y & 0 & 0 \\ -\alpha'_y\Delta c_x + c'_z & 0 & 0 \end{bmatrix}, \\ \mathcal{T}_2 &= \begin{bmatrix} -\alpha''_z c'_x + \alpha'_z c''_x & (1 + \alpha'_f + \alpha''_f)c'_x & \alpha''_x c'_x \\ -(1 + \alpha'_f + \alpha''_f + \alpha''_r)c''_x & \alpha'_z c'_x - \alpha''_z c''_x & \alpha''_y c''_x \\ -\alpha'_x c''_x & -\alpha'_y c'_x & 0 \end{bmatrix}, \\ \mathcal{T}_3 &= \begin{bmatrix} (\alpha''_y + p''_x)c'_x - \alpha'_y c''_x & -\alpha''_x c'_x & (1 + \alpha'_f)c'_x \\ \alpha'_x c''_x & 0 & \alpha'_z c'_x \\ -(1 + \alpha''_f + \alpha''_r)c''_x & -\alpha''_z c''_x & -\alpha'_y c'_x + \alpha''_y c''_x \end{bmatrix}, \end{aligned} \quad (3.30)$$

where the following substitution was performed in order to increase the readability: $\Delta c_x = (c'_x - c''_x)$.

3.3.3 Solving Set of Linear Equations

Given a set of triplet point matches $[u, v, u', v', u'', v'']$, the aim is now to find the unknown variables numerically. Special attention is needed to estimate the components c'_x and c''_x . In fact, what is required for a multi-baseline stereo approach is the ratio of the baselines between the three cameras. The baseline between the first and the second camera can be normalized: $\beta_{12} = 1$. The relative size of the baseline between first and third camera is $\beta_{13} = c''_x/c'_x$. It can also be expressed as sum of the baselines between first and second camera on one hand, and the baseline between the second and the third cameras on the other, i.e.: $\beta_{13} = \beta_{12} + \beta_{23}$. The baseline ratios β_{23} and β_{13} can't be estimated linearly in a single step as they form products with other variables. Consequently, they need to be

estimated iteratively. Given a start value for $\beta_{13} = \beta_{12}$ the value is updated with each iteration by adding a correction value α_b according to the following equation:

$$c_x''/c_x' = \beta_{13} + \alpha_b \quad (3.31)$$

Given the trifocal tensor in slices representation and a set of triplet point matches $[u, v, u', v', u'', v'']$ a set of nine equations can be derived by inserting eqns. (3.30) and (3.31) into eqn. (2.40). As an example the equation corresponding to the 2nd row and column of eqn. (2.40) is presented:

$$\begin{aligned} (u - u')\alpha_b + (\beta_{23}u + u'')\alpha_f' + (\beta_{23}u - \beta_{13}u')\alpha_f'' + (\beta_{23}u - \beta_{13}u')\alpha_r'' - (\beta_{13}u'v)\alpha_x' \\ + (u''v)\alpha_x'' + (\beta_{23}uu' + u'u'' + \beta_{13})\alpha_y' \\ + (\beta_{23}u u'' - \beta_{13}u'u'' - 1)\alpha_y'' - (\beta_{13}v)\alpha_z' + v\alpha_z'' - p_x'' + \beta_{23}u - \beta_{13}u' + u'' = 0 \end{aligned} \quad (3.32)$$

The coefficients can be grouped to formulate a constraint matrix \mathbf{A} forming a linear set of equations which allows finding the vector \mathbf{x} of geometric parameters:

$$\mathbf{x} = [\alpha_b, \alpha_f'', \alpha_f', \alpha_r'', \alpha_x', \alpha_x'', \alpha_y', \alpha_y'', \alpha_z', \alpha_z'', p_x'']^T. \quad (3.33)$$

The matrix \mathbf{A} consists of eleven columns and nine rows per point triplet. The terms of the equation which do not depend on a variable can be grouped on the right side of the equation forming the vector \mathbf{b} . Finally, the linear set of equations can be solved by QR-factorization [Bronstein95].

$$\mathbf{Ax} = \mathbf{b}. \quad (3.34)$$

3.3.3.1 Iterative Estimation

The precision of the geometric parameters can be improved by iterating the estimation process, i.e. the rectifying homographies are applied to the feature point triplets and a new Taylor expansion is performed. The process is iterated until no further improvement is achieved, i.e. if the back-projection error [Hartley04] for the resulting rectified projection matrices converges.

3.3.4 Trifocal Rectification

Once, the geometric parameters from eqn. (3.33) are known, it is easy to derive rectifying homographies. Related work from the field of multi-camera rectification was presented in section 2.4.3.1. Unlike the approaches of the related work, the method presented in this chapter is able to ensure a horizontal alignment as well as a vertical alignment.

The rectifying homographies are composed using the geometric parameters obtained in eqn. (3.34). The homographies shall transform the projection matrices \mathbf{P}' and \mathbf{P}'' such that they have the same orientation and intrinsic parameters as the reference camera \mathbf{P} . Thus, given the matrices

$$\mathbf{P}' = \mathbf{K}'\mathbf{R}'[\mathbf{I}|\mathbf{-C}'_x], \mathbf{P}'' = \mathbf{K}''\mathbf{R}''[\mathbf{I}|\mathbf{-C}''_x] \quad (3.35)$$

the following homographies \mathbf{H}' and \mathbf{H}'' can be formulated:

$$\begin{aligned} \mathbf{H}' &= \mathbf{K}\mathbf{R}'(\alpha'_x, \alpha'_y, \alpha'_z)^{-1} \mathbf{K}'(\alpha'_f)^{-1} \\ \mathbf{H}'' &= \mathbf{K}\mathbf{R}''(\alpha''_x, \alpha''_y, \alpha''_z)^{-1} \mathbf{K}''(\alpha''_f, \alpha''_r, p''_x)^{-1} \end{aligned} \quad (3.36)$$

where $\mathbf{R}'(\alpha'_x, \alpha'_y, \alpha'_z)$ and $\mathbf{R}''(\alpha''_x, \alpha''_y, \alpha''_z)$ denote rotation matrices around the x -, y -, and z -axis respectively according to eqn. (2.18). After applying the homographies, the projection matrices from eqn. (3.37) have the desired simplified form corresponding to an ideal linear camera array.

Algorithm 3.1. Iterative Estimation of the Trifocal Tensor

Input : triplet point matches $[u, v, u', v', u'', v'']$

Output : Final geometric parameter vector \mathbf{x}_f
 Baseline ratios $\beta_{12}, \beta_{23}, \beta_{13}$
 Projection Matrices \mathbf{P}' and \mathbf{P}''
 Rectifying Homographies \mathbf{H}' and \mathbf{H}''

Initialize parameter vector \mathbf{x}_f and baseline ratios

$\mathbf{x}_f = \vec{\mathbf{0}}; \beta_{12} = 1; \beta_{13} = \beta_{12}; \beta_{23} = 0;$

$\mathbf{m}_0 = [u, v, u', v', u'', v'']; \mathbf{m}_i = \mathbf{m}_0; \mathbf{err}_{bp} = \max$

do Iterate until back-projection error converges

do Iterate until correction value α_b vanishes

 build new constraint matrix \mathbf{A} and solve eqn. (3.34)

$[\mathbf{A}, \mathbf{b}] = \text{get_constraint_matrix}(\mathbf{x}_f, \beta_{13}, \beta_{23}, \mathbf{m}_i)$

$\mathbf{x}_i = \text{solve_linear_system}(\mathbf{A}, \mathbf{b})$

$\beta_{13} += \mathbf{x}_i["\alpha_b"]; \beta_{23} = \beta_{13} - \beta_{12}$

until $|\mathbf{x}_i["\alpha_b"]| < \epsilon$

$\mathbf{x}_f += \mathbf{x}_i;$ Add parameter vector from last iteration

 Projection matrices according to eq. (3.35)

$[\mathbf{P}', \mathbf{P}''] = \text{get_proj_matrices}(\mathbf{x}_f)$

 Homographies according to eq. (3.36)

$[\mathbf{H}', \mathbf{H}''] = \text{get_homographies}(\mathbf{x}_f)$

 Apply homographies to feature points

$\mathbf{m}_i = \text{get_rectified_points}(\mathbf{m}_0, \mathbf{H}', \mathbf{H}'')$

 Retrieve back-projection error after Rectification, eq. (3.37)

$\mathbf{err}_{old} = \mathbf{err}_{bp}; \mathbf{err}_{bp} = \text{back_proj_err}(\mathbf{m}_i, \mathbf{H}'\mathbf{P}', \mathbf{H}''\mathbf{P}'')$

until $|\mathbf{err}_b - \mathbf{err}_{old}| < \epsilon$

For a better understanding, the iterative estimation of the trifocal tensor is summarized as pseudo-code in Algorithm 3.1.

$$\mathbf{H}'\mathbf{P}' = \begin{bmatrix} 1 & 0 & 0 & c'_x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{H}''\mathbf{P}'' = \begin{bmatrix} 1 & 0 & 0 & c''_x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \quad (3.37)$$

Please note that as the reference camera \mathbf{P} remains unchanged during the rectification process, i.e. it can serve as reference camera for multiple camera triplets within a linear camera array.

3.3.5 Results

In a first experiment, the proposed algorithm was tested using synthetic data. Varying noise amplitudes of the feature point positions and three different mechanical alignment qualities were

simulated with subsequent analysis of the back-projection error and baseline ratios. For each angle $(\alpha'_x, \alpha''_x, \alpha'_y, \alpha''_y, \alpha'_z, \alpha''_z)$ a deviation of 1° (5°) was used to simulate low (high) rotation errors. The intrinsic parameters $(\alpha'_f, \alpha'_r, \alpha''_f, \alpha''_r, p'_x)$ were set to $f/100$ ($5 \times f/100$) to simulate low (high) intrinsic errors. The camera centers were moved in y - and z -direction by 1% (5%) of the baseline c'_x to simulate low (high) translational errors. As shown in Table 3.3, the resulting back-projection errors are in the range of the noise amplitude and increase with higher mechanical alignment errors.

Table 3.3. Back-projection error err_{bp} estimated using synthetic data with increasing noise level and alignment errors. Beside the back-projection error, the baseline ratio β_{13} is estimated. The ground truth value for β_{13} is 5.0.

Noise level σ_n (in pixel)	Mechanical Alignment Quality		
	Good low rot. & intrinsic err.	Med. low rot., intr. & transl. err.	Bad high rot., intr. & transl. err.
	err_{bp}/β_{13}	err_{bp}/β_{13}	err_{bp}/β_{13}
0.0	0.000 / 5.000	0.311 / 5.000	1.554 / 4.989
0.1	0.086 / 5.000	0.331 / 5.000	1.558 / 4.989
0.2	0.172 / 5.000	0.371 / 4.999	1.571 / 4.989
0.5	0.431 / 4.999	0.548 / 4.998	1.646 / 4.988
1.0	0.864 / 4.996	0.924 / 4.994	1.846 / 4.987
2.0	1.733 / 4.983	1.765 / 4.981	2.439 / 4.970
5.0	4.370 / 4.892	4.382 / 4.902	4.768 / 4.869

In Figure 3.5, the rectification quality of synthetic image data is shown with a noise level σ_n of 0.5 pixels and medium mechanical alignment quality.

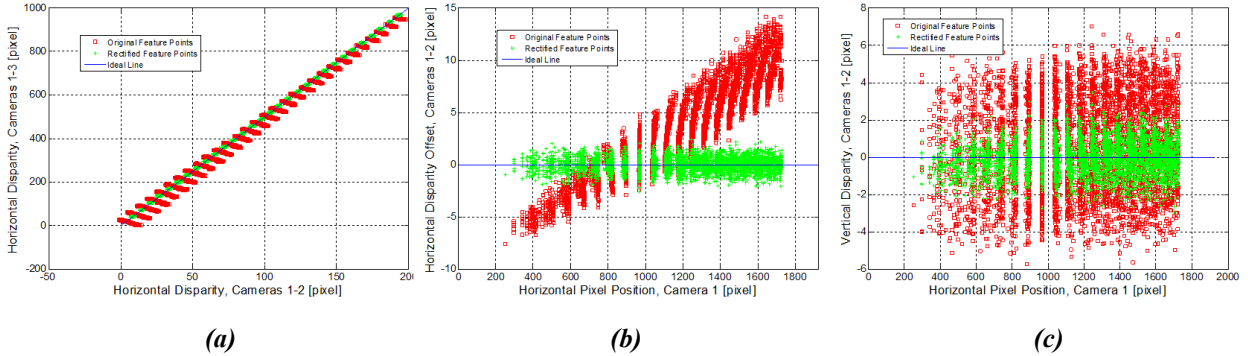


Figure 3.5. Evaluation of the rectification quality using synthetic data with a noise level of $\sigma_n = 0.5$ pixels and a medium mechanical alignment quality according to Table 3.3. The back-projection error ($err_{bp} = 0.548$) and the baseline ratios $\beta_{12} = 1.0$, $\beta_{23} = 3.998$, and $\beta_{13} = 4.998$ were determined according to the proposed estimation algorithm. (a) The horizontal disparities of the unrectified synthetic data (original feature points marked in red) are not proportional which is the case instead for the rectified points (green), which lie mainly on the ideal blue line. (b) The horizontal disparities of the original feature points (red) have a considerable offset from the ideal position, e.g. proportionality. After rectification (green points), the offset has been minimized. (c) The original feature points (red) show vertical disparities. After rectification, the vertical disparities have been minimized.

In a second experiment the multi-camera footage was used originating from the MUSCADE project [Muscade]. It was shot using high quality HD cameras with mixed stereo baseline (cf. chapter 6 for details). For the experiment described here, the three left-most cameras were used. Figure 3.6 (top) shows a rectified image triplet along with horizontal lines demonstrating that corresponding pixels lie

on the same image scanline. Moreover, the disparity maps in inverse-of-a-distance representation [Okutomi93] in Figure 3.6 (bottom) demonstrate the quality of the horizontal alignment (cf. also section 6.4). In Figure 3.7 the rectification accuracy is illustrated by analyzing the vertical and horizontal alignment before and after rectification of feature points estimated using the approach described in chapter 4 involving the feature descriptor SKB.



Figure 3.6. Rectified multi-camera footage (top) along with normalized disparity maps (bottom). Horizontal reference lines illustrate that the vertical disparities vanished. The disparity maps were normalized using the baseline ratios β_{12} , β_{23} , and β_{13} resulting in corresponding gray scale values for corresponding pixels despite non-equidistant baselines.

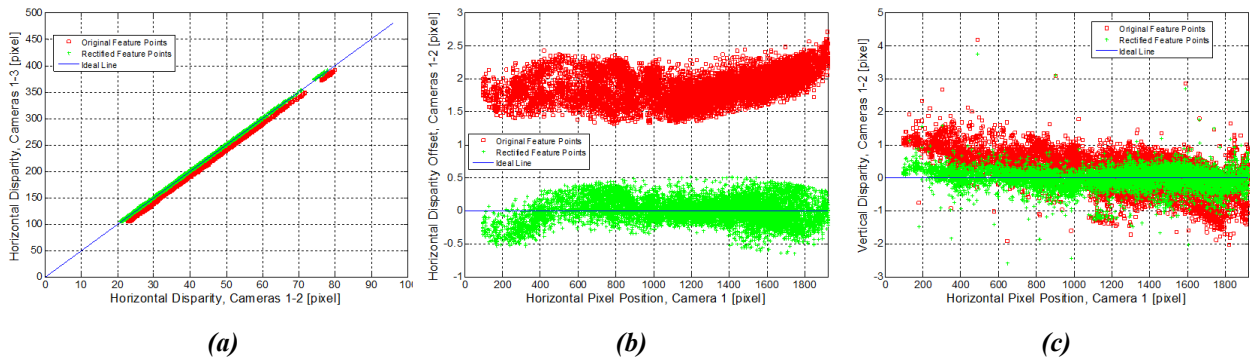


Figure 3.7. Evaluation of the rectification quality for the multi-camera footage from Figure 3.6. The back-projection error ($\text{err}_{bp} = 0.298$) and the baseline ratios ($\beta_{12} = 1.0$, $\beta_{23} = 4.004$, and $\beta_{13} = 5.004$) were determined according to the proposed algorithm for the linearized estimation of the trifocal tensor. (a) The horizontal disparities of the original feature points (red) are not proportional which is the case for the rectified points (green), which lie mainly on the ideal blue line. (b) The horizontal disparities of the original feature points (red) have a considerable offset from the ideal position, e.g. proportionality. After rectification (green points), the offset has been minimized. (c) The original feature points (red) show vertical disparities. After rectification, the vertical disparities have been minimized.

3.4 Conclusion

A method to linearize projective entities has been proposed in this chapter. The concept was applied to two main applications, the estimation of a linearized fundamental matrix and the estimation of a linearized trifocal tensor.

Along with the linearized fundamental matrix a new rectification technique which allows computing rectifying homographies was proposed. The algorithm was combined with a RANSAC elimination of outliers. The algorithm uses point correspondences and does not need prior knowledge of the projection matrices. The technique involves a linearized computation of the epipolar geometry which makes it suitable for setups which are near the rectified state. The image distortions induced by the rectification have shown to be negligible compared to techniques proposed in [Loop99], [Hartley99], and [Mallon05] and also slightly reduced compared to [Georgiev13]. The algorithm preserves the convergence plane of the stereo setup and is therefore suitable for rectifying 3D-TV stereo sequences [Woods93]. These are important requirements for the target application described in chapter 5, i.e. a stereoscopic assistance system.

The linearized trifocal tensor and its estimation were adapted to the geometric properties of a linear camera array. Similar to the two-camera case, rectifying homographies were estimated jointly with the linearized trifocal tensor. The multi-camera rectification results show that the method based on feature point triplets from uncalibrated cameras is able to perform a vertical and horizontal alignment ensuring proportional horizontal disparities. This makes the algorithm suitable for the multi-camera disparity estimation algorithm described in chapter 6.

4 Fast Feature Point Description and Matching

4.1 Introductory Remarks

Within this chapter a new approach for feature description used in image processing and robust image recognition algorithms such as 3D camera tracking, view reconstruction or 3D scene analysis is presented. The main contribution is a new feature descriptor called **Semantic Kernels Binarized (SKB)**¹⁵.

An overview of related work regarding feature detection, description, and matching was given in section 2.5. Compared to the state-of-the-art, the BRIEF descriptor [Calonder10] is nearest to the method proposed in this chapter. Consequently, the proposed descriptor will be compared to the BRIEF-256 descriptor in section 4.6.2. A comparison to other state-of-the-art descriptors is given in section 4.6.1.

The SKB descriptor as described in [Zilly11c] has been used by Stefanoski et al. [Stefanoski13] for an implementation of the proposed image domain warping algorithm along with a comparison of the SKB with respect to its suitability for the algorithms proposed by Stefanoski et al.

4.2 Basic Properties of the SKB-Descriptor

The basic idea of the SKB descriptor proposed in this chapter is that several convolutions with a set of folding kernels are performed on the support region. The kernels consist of basic geometric structures, namely edges, ridges, corners, blobs, and saddles. These kernels have a dedicated meaning in computer vision, thus the naming *semantic kernels*. As the filter responses are binarized, the descriptor is denominated as *semantic kernels binarized*, or SKB.

The major contribution within this chapter is a new robust descriptor. In total 16 different kernel responses are evaluated at 16 positions within the support region. The 16 kernels represent basic geometric structures as mentioned above. This results in 256 dimensions of the descriptor. However, the filter responses can be binarized. Consequently, despite the high descriptor dimension, only 256 bit, i.e. 32 bytes are needed. A second variant of the proposed descriptor uses 2 bits per dimensions, which results in 64 bytes per keypoint. A low memory usage per keypoint is particularly beneficial, if many descriptors need to be stored or transmitted over a network with limited bandwidth. As comparison, SIFT uses 128 floating point numbers, resulting in a descriptor size of $4 \times 128 = 512$ bytes. Binarization of the filter responses has an additional advantage. As described in section 2.5.4, interest point matching can be implemented in a very efficient way in the case of binarized descriptors such as BRIEF [Calonder10]. Optimized SSE 4.2 operations can be performed in order to evaluate the

¹⁵ Parts of the content in this chapter have been previously published in [Zilly11c].

Hamming distance between two descriptors. The design of the descriptor allows an efficient implementation in hardware. An FPGA implementation of the SKB is described in [Schaffner13].

4.3 Defining the Support Region

The first step of the description process is to interpolate the support region based on the subpixel position and scale of the interest point. If required, one can achieve rotational invariance by turning the support region such that the main gradient direction points northwards. If feature points shall be matched within two nearly rectified cameras, rotation invariance can be omitted.

Two types of support regions will be proposed which address two different types of interest point detectors:

Type A: support region consisting of 12x12 pixels and overlapping kernel set evaluation regions, optimized for interest points with complex gradient structure in the center of the support region, e.g. corner detectors.

Type B: support region of 16x16 pixels with equidistant and non-overlapping kernel set evaluation regions, optimized for interest points with uniform gradient structure, or interest points where most of the gradients occur at the border of the support region which is the case when using blob detectors.

4.3.1 Overlapping Kernel Set Evaluation Regions (Type A)

The process of interest point description will now be described in more detail. It is assumed that an interest point has been detected and a normalized support region of 12x12 pixels around the interest point has been computed.

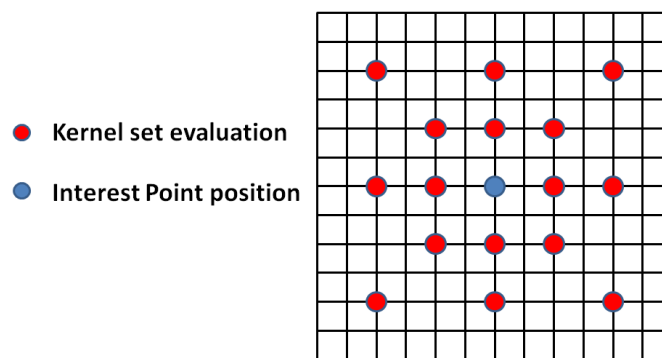


Figure 4.1. All kernels are evaluated at the positions around the red circles. The blue circle indicates the position of the interest point. The kernels have a size of 4x4 pixels, such that their respective evaluation regions overlap.

Around each of the red dots in Figure 4.1, a number of 16 kernels will be evaluated. An example is shown in Figure 4.2. All 16 kernels are shown in Figure 4.5 to Figure 4.8.

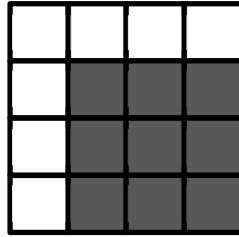


Figure 4.2. Example of a corner-like kernel of size 4x4.

Apparently, the regions of evaluation overlap as the kernels have a size of 4x4 and some evaluation points are only two pixels apart. Some pixels contribute to one, two, or three evaluation points. Figure 4.3 illustrates this behavior. The inner region will be evaluated three times, the outer pixel only once. This approximates a Gaussian giving more statistical weight to the center pixels.

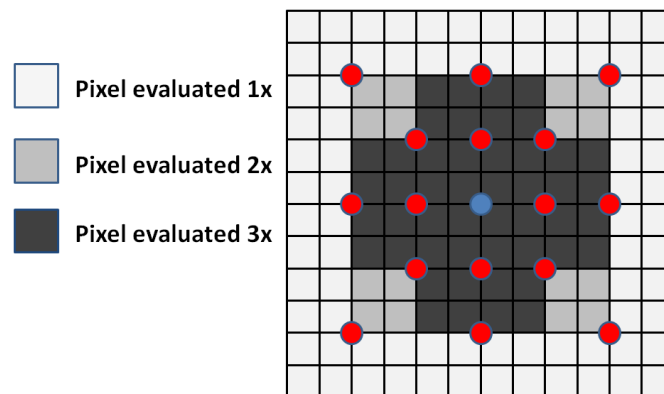


Figure 4.3. Overlap of the kernel set evaluation regions. The inner region will be evaluated three times, the outer pixels only once. This approximates a Gaussian giving more statistical weight to the center pixels.

This type of description is suitable when the most important gradients are near the interest point. This is the case for region detectors or interest point detectors which search for corners, saddles, etc. In total, 256 filter responses are calculated.

4.3.2 Uniform Kernel Set Evaluation Regions (Type B)

In the case of blob detectors, one should use the version with a support region of 16x16 pixels. The sampling is now uniform and non-overlapping. Figure 4.4 illustrates this type of support region. All 16 kernels are evaluated at the red dot positions. The kernels have a size of 4x4 pixels, thus, the evaluated regions do not overlap.

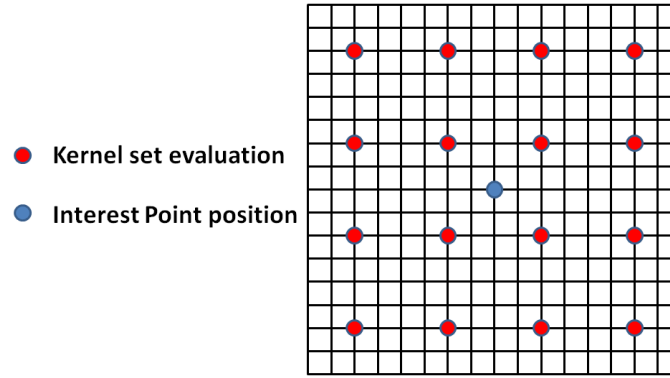


Figure 4.4. Support region suitable for blob detectors. Equidistant and non-overlapping regions for kernel set evaluation.

Blobs do not have a complex gradient structure at the center. The important gradients lie at the border of the support region. This type of support region is more suitable when the descriptor is combined with a blob-based interest point (or region) detector. The principles of blob detectors are described in section 2.5.2.2. Just like in the *type A* version, 16 kernels are evaluated at 16 positions resulting in 256 filter responses.

4.4 The Kernels

The support regions are convoluted using 16 different kernels. All kernels represent a named geometric structure justifying the terminology of *semantic kernels*. Many kernels exist for different orientations. For instance, four main directions are evaluated among the edge-like kernels. All kernels are illustrated in Figure 4.5 to Figure 4.8.

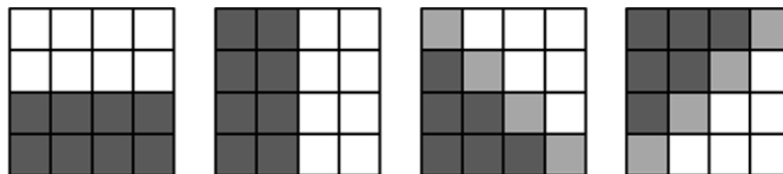


Figure 4.5. Edge-like kernels. Four different main directions are evaluated.

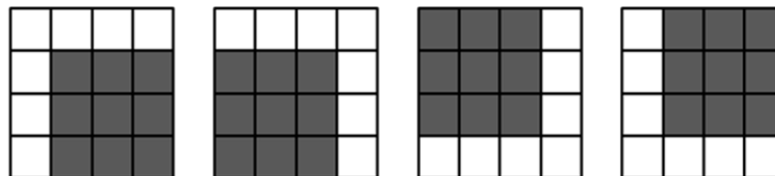


Figure 4.6. Corner-like kernels.

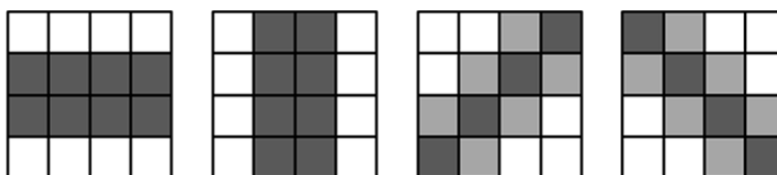


Figure 4.7. Ridge-like kernels.

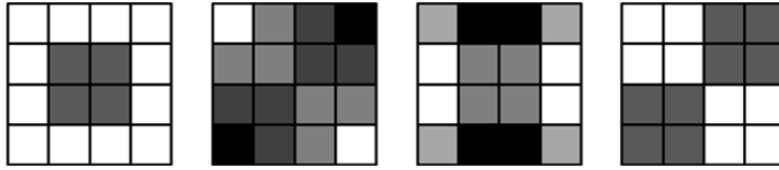


Figure 4.8. Blob-like and saddle-like kernels.

4.4.1 Filter Response Binarization

The filter responses resulting from the convolution with the semantic kernels can be binarized. As mentioned, 16 kernels are evaluated at 16 positions of the support region.

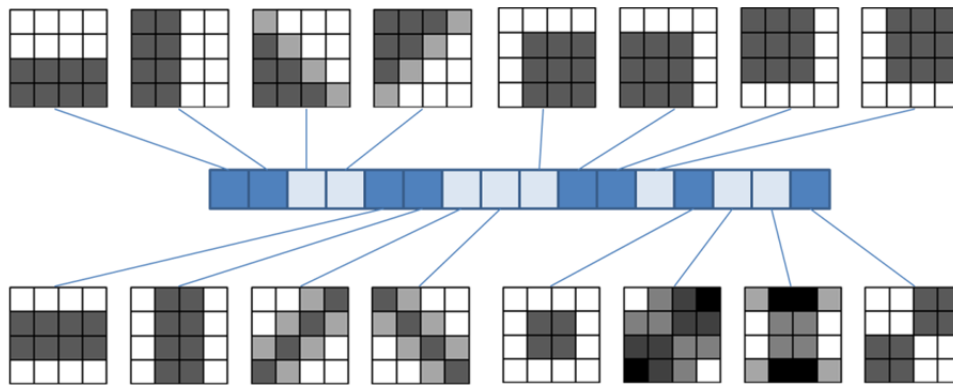


Figure 4.9. For each kernel and evaluation position, the result of the folding operation is stored in the resulting key point descriptor.

The concept is illustrated in Figure 4.9 and Figure 4.10. The length of the feature descriptor depends on the binarization strategy. In the example from Figure 4.10 a single bit is used to store the filter response in the feature vector describing the key point.

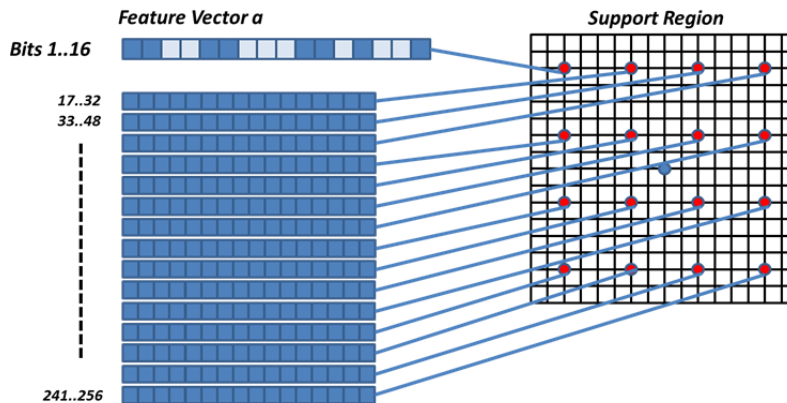


Figure 4.10. For each evaluation point, all 16 kernels are evaluated. In the simplest case, each filter response leads to a single bit within the feature vector describing the key point (variant A is illustrated). The principle can be applied to both versions of support region. In this example, the type B support region suitable for blob detectors is shown.

However, it is possible to spend two or more bits per filter response. There are three proposed variants for the SKB to perform this task, namely variant A resulting in 256 bits and the variants B and C with 512 bits per key point.

Variant A: 256 Bits

In the easiest case only the signum of the filter response is used. In that case, the descriptor has a dimension of 256 bits. It is not normalized, i.e. the number of bits which are set can differ from key point to key point. Matching can later be performed by evaluating the Hamming distance between two descriptors.

Variant B and C: 512 Bits

Variants B and C use 2 bits per filter response r to describe a tri-state situation. Three cases are distinguished: positive answer of high amplitude, positive or negative answer of low amplitude, and negative answer of high amplitude. The resulting bit pair b depends on the filter response r and the threshold θ_{low} which discriminates between low and high:

$$b = \begin{cases} 01: & r > \theta_{low} \\ 00: & -\theta_{low} \leq r \leq \theta_{low} \\ 10: & r < -\theta_{low} \end{cases} \quad (4.1)$$

In variant B, the threshold θ_{low} is fixed. This makes the computation faster, but also impairs the robustness against contrast changes as the descriptor is not normalized. In order to conduct the matching of the key points, the Hamming distance needs to be evaluated.

In variant C, the threshold θ_{low} is self-normalized. The aim is to normalize the number of bits set. A unit length u is defined to normalize the descriptor. The threshold θ_{low} is now chosen such that for each descriptor, exactly u bits are set. The matching process can later be performed by evaluating the scalar product.

As the descriptor comprises 512 bits, representing a tri-state, the unit length u is calculated as following:

$$u = \left\lceil \frac{512}{3} \right\rceil = 171. \quad (4.2)$$

4.5 Matching Approach

One of the main advantages of the proposed descriptor is the fast matching ability. The descriptors are binary numbers. Consequently the scalar product which is needed for a correlation matrix can be computed very efficiently as well as the Hamming distance between two descriptors.

Only eight executions of this instruction are needed for the 512 dimensional feature vector (variant C) and only 4 executions are needed for a 256 dimensional feature vector (variants A and B).

4.5.1 Additional Matching Constraints

The best matching feature vector (variant B and C: lowest Hamming Distance, variant A: highest scalar product) is compared to the second best. To achieve better matching results, it is demanded that the scalar product of the best match is s times higher than the second best match. If the Hamming distance is evaluated, it is demanded that the lowest distance is s times lower than the second lowest distance. This ensures that the matching feature points are unique. If the second best match had a similar matching score, the match would be less robust as noise within the image pair would have a too high influence, e.g. the same scene shot at a different moment might lead to another best match.

4.6 Evaluation of SKB

Different comparison tests have been conducted. In a first step the three variants of the SKB descriptor were compared using the evaluation framework and dataset used in [Mikolajczyk05a]¹⁶. The test images in the dataset are affected by image degradation (blur, illumination change, Jpeg compression) at six levels of degradation within each set.

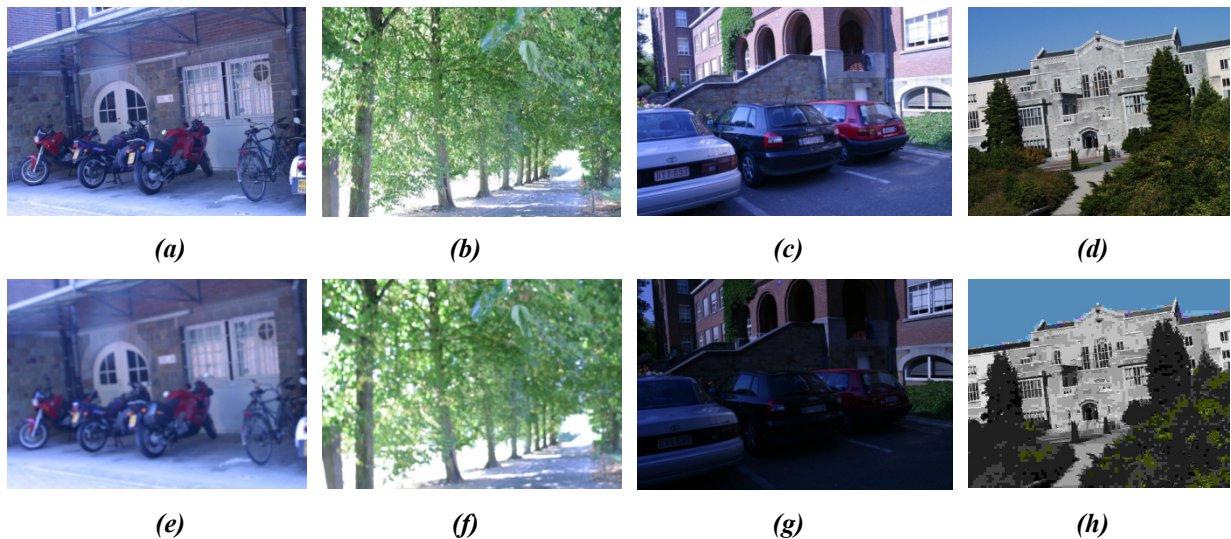


Figure 4.11. Test images from the evaluation framework [Mikolajczyk05a]. In the top row, the 1st images from the respective image series are shown which are not affected by image degradation. In the bottom row the 6th image from the respective series are shown which show the strongest artefacts. (a) Bikes: image 1. (b) Trees: image 1. (c) Leuven: image 1. (d) UBC: image 1. (e) Bikes: image 6 (blur). (f) Trees: image 6 (blur). (g) Leuven: image 6 (illumination). (h) UBC: image 6 (Jpeg compression).

Subsequently, the fastest SKB variant was compared with the BRIEF-256 descriptor [Calonder10] to compare the recall rate and the processing time of the description process. In a last test the runtime of a stereo matching application and its detection quality in a real-time environment was evaluated. A comparison of the SKB regarding its suitability for the image domain warping algorithm is performed in [Stefanoski13]. Another comparison of different variants of the SKB was conducted in [Schaffner13].

¹⁶ The data set is available at <http://www.robots.ox.ac.uk/~vgg/research/affine>.

4.6.1 SKB vs. GLOH, SIFT and Cross Correlation

In a first test the three variants of the SKB descriptor were compared with rotationally invariant versions of GLOH and SIFT, and a standard cross correlation as defined in eqn. (2.59) using the test framework provided in [Mikolajczyk05a] as shown in Figure 4.11. As region detector the Hessian-Laplace detector was chosen. This detector is part of the region detectors provided in the test framework. Four images from the data set were chosen which do not require rotational invariance. The latter is not necessary in the case of nearly rectified stereo images, a constraint which is met by the target application, i.e. stereoscopic image analysis as described in chapter 5. Figure 4.12 to Figure 4.15 show the results of this performance evaluation. For all test images, image 1 has been matched with the respective image 6, which shows the strongest artifacts, i.e. Blur for the *Bikes* and *Trees* images, Jpeg compression for the *UBC* image, and illumination change for the *Leuven* image. The descriptor shows a very high precision when the threshold is chosen to reproduce a recall rate around 50% of the maximum recall rate. By using the terms recall rate and “1-precision” as defined in eqns. (2.56) and (2.57), the terminology proposed by [Ke04] is followed. For applications such as stereo matching and calibration purposes, it is important to have a low outlier rate while a reasonable number of feature points is sufficient for this task. In other words, the aim was not to maximize the recall for high 1-precision values, but to maximize the precision at a given recall rate, given that the feature points shall be suitable for applications such as the estimation of a fundamental matrix using RANSAC fitting as described in section 3.2.5. Given that the value of 1-precision from Figure 4.12 to Figure 4.15 corresponds to the outlier rate ϵ from eqn. (3.16) one can see that a low outlier rate is favorable for the RANSAC process.

For the evaluation described here, 50% of the maximum recall rate was chosen as target value. One can observe that the three variants A, B and C show strong results at low thresholds for the nearest neighbour matching. In that scenario, the precision is high compared to other descriptors (*Bikes*, *UBC*, *Trees*). However, in the test image *Leuven*, which tests for strong illumination changes, other descriptors show a better performance.

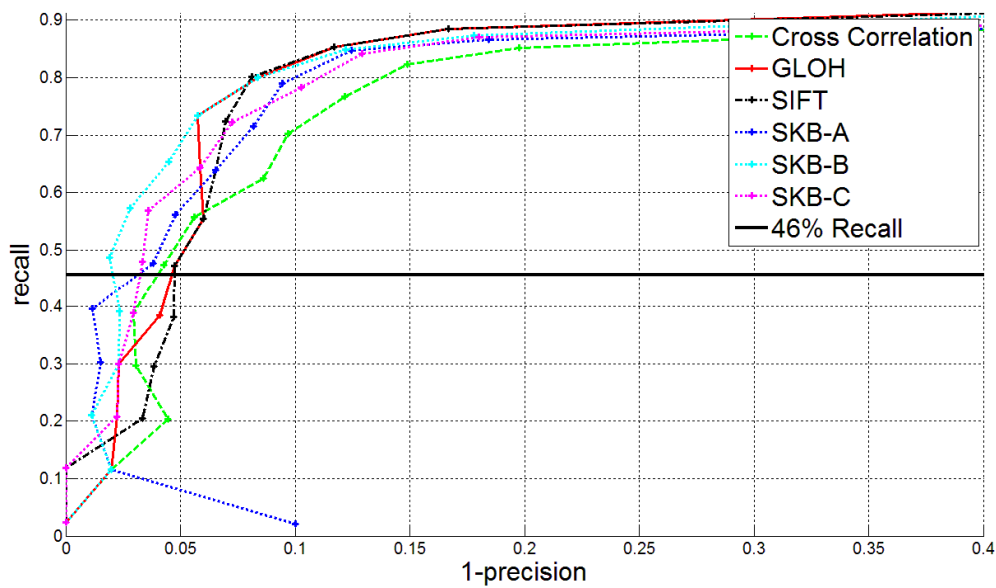


Figure 4.12. Results for image “Bikes” (affected by image blur). The three variants SKB-A, SKB-B, and SKB-C have a similar performance. They show a particularly good performance at low thresholds compared to SIFT, GLOH, and Cross Correlation. At a recall rate of 46% which is half of the maximum recall rate, SKB-B has the highest precision, i.e. the fewest outliers.

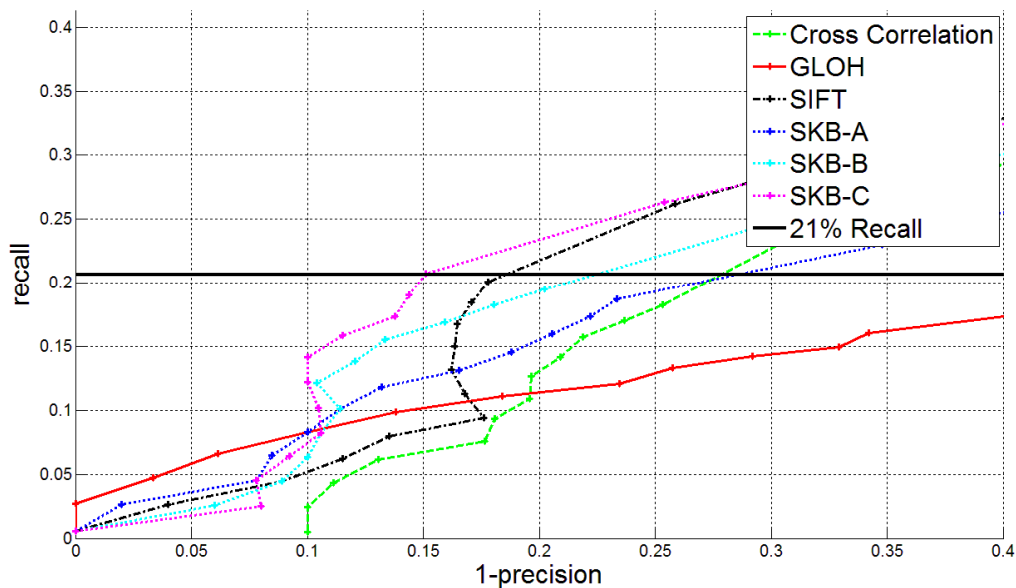


Figure 4.13. Results for image “Trees” (affected by image blur). The three variants SKB-A, SKB-B, and SKB-C have a similar performance. They show a particularly good performance at low thresholds compared to SIFT, GLOH, and Cross Correlation. At a recall rate of 21% which is half of the maximum recall rate, SKB-C has the highest precision, i.e. the fewest outliers.

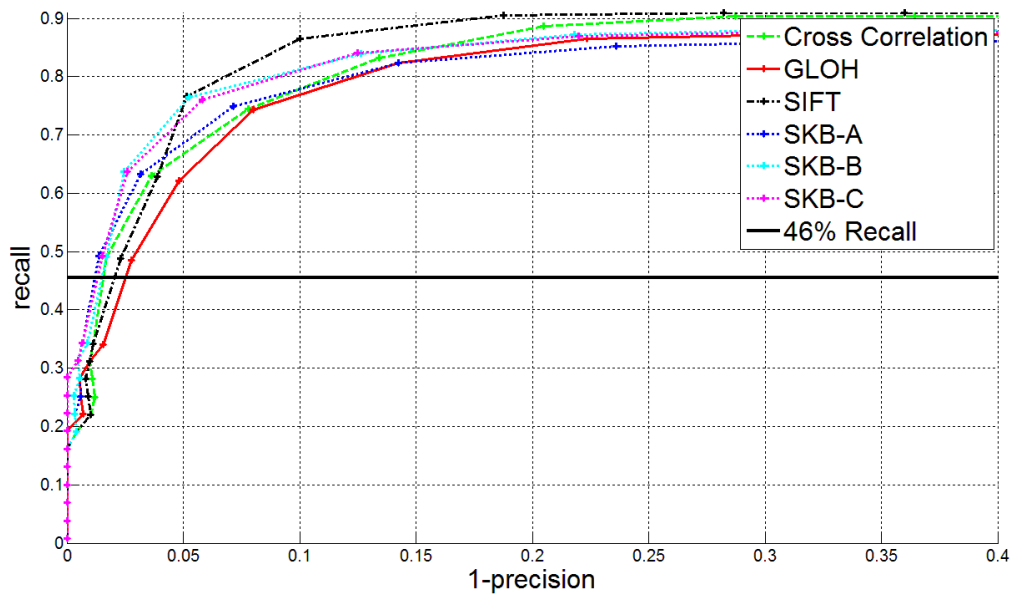


Figure 4.14. Results for image “UBC” (Jpeg compression artifacts). The three variants SKB-A, SKB-B, and SKB-C have a similar performance. They show a particularly good performance at low thresholds compared to SIFT, GLOH, and Cross Correlation. At a recall rate of 46% which is half of the maximum recall rate, SKB-A has the highest precision, i.e. the fewest outliers.

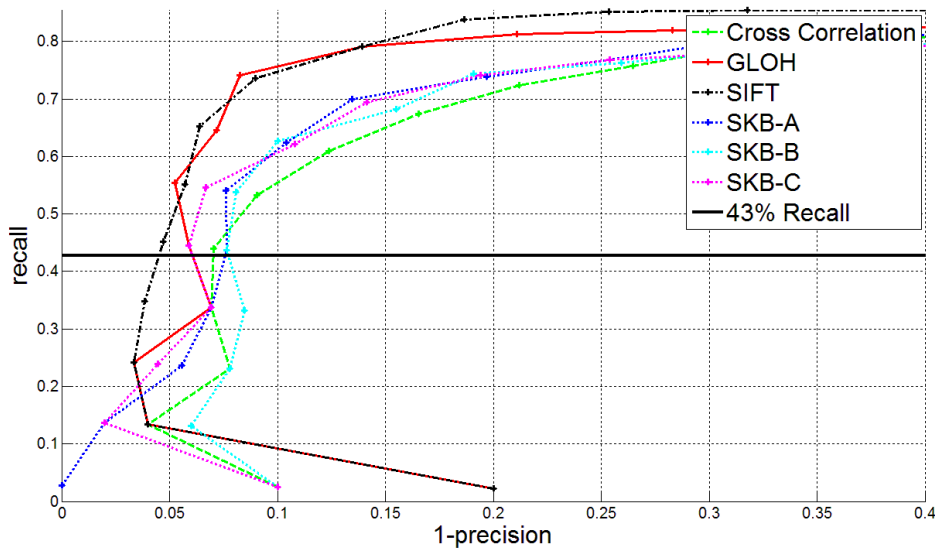


Figure 4.15. Results for image “Leuven” (Illumination change). The three variants SKB-A, SKB-B, and SKB-C have a similar performance. They perform better than Cross Correlation, but worse than SIFT and GLOH. At a recall rate of 43% which is half of the maximum recall rate, SIFT has the highest precision, i.e. the fewest outliers. GLOH and SKB-C have the same precision at that recall rate.

4.6.2 SKB vs. BRIEF

The fastest SKB variant (SKB-A) was compared to the OpenCV 2.2 implementation of BRIEF-256 as described in [Calonder10]. Both descriptors use 256 bit descriptors and can be matched using a Hamming distance matcher. A test was performed where the recall rate was calculated for both descriptors, again using the test images provided in [Mikolajczyk05a]. The test data set, as mentioned

above, contains images that are affected by different types of degradation at six levels of degradation within each set.

For different image pairs, the SURF detector [Bay08] was used to detect interest points for the first image. Subsequently, this interest point was transferred using a ground truth homography into the second image. This approach is also performed in [Calonder10]. The homographies and images are part of the test framework provided in [Mikolajczyk05a]. In the next step, the interest points were described using SKB-A and BRIEF-256. In a last step, the feature points were matched using a nearest neighbor matcher. Due to the ground truth homography, one can examine the rate of correct and false matches.

For each image set, image 1 is subsequently matched against the images 2, 3, 4, 5, and 6. The image 1 in each set is the version without degradation while the amount of image degradation increases subsequently with a maximum degradation for image 6. Consequently, the recall rate for the comparison between image 1 and 2 is higher than the comparison between images 1 and 6. The result is illustrated in Figure 4.16. As one can see, the recall rate is higher for SKB in all cases.

Finally, the runtime of SKB-A, SIFT, SURF, and BRIEF-256 descriptors were compared when describing 8372 feature points. For SIFT, SURF and BRIEF the OpenCV 2.2 implementation was used, running on a single core CPU at 3.33 GHz. The matching process itself was not compared, as it is identical between BRIEF-256 and SKB-A.

Table 4.1. Runtime of different descriptors. 8375 feature points needed to be described. SKB-A performs the fastest even when the calculation of the integral image is included. Without counting the time for the integral image, it runs around 2.5 times faster than BRIEF-256.

<i>Descriptor</i>	<i>Runtime in ms</i>
SIFT	2959
SURF	457
BRIEF-256	42
SKB-A (incl. Integral Image)	24
SKB-A	17

Table 4.1 shows the result of the runtime comparison. SKB-A performs the fastest with 24 ms for 8372 feature points, including the time required for the computation of the integral image. If the integral image is already available (e.g. already computed by the interest point detector) SKB-A needs only 17 ms.

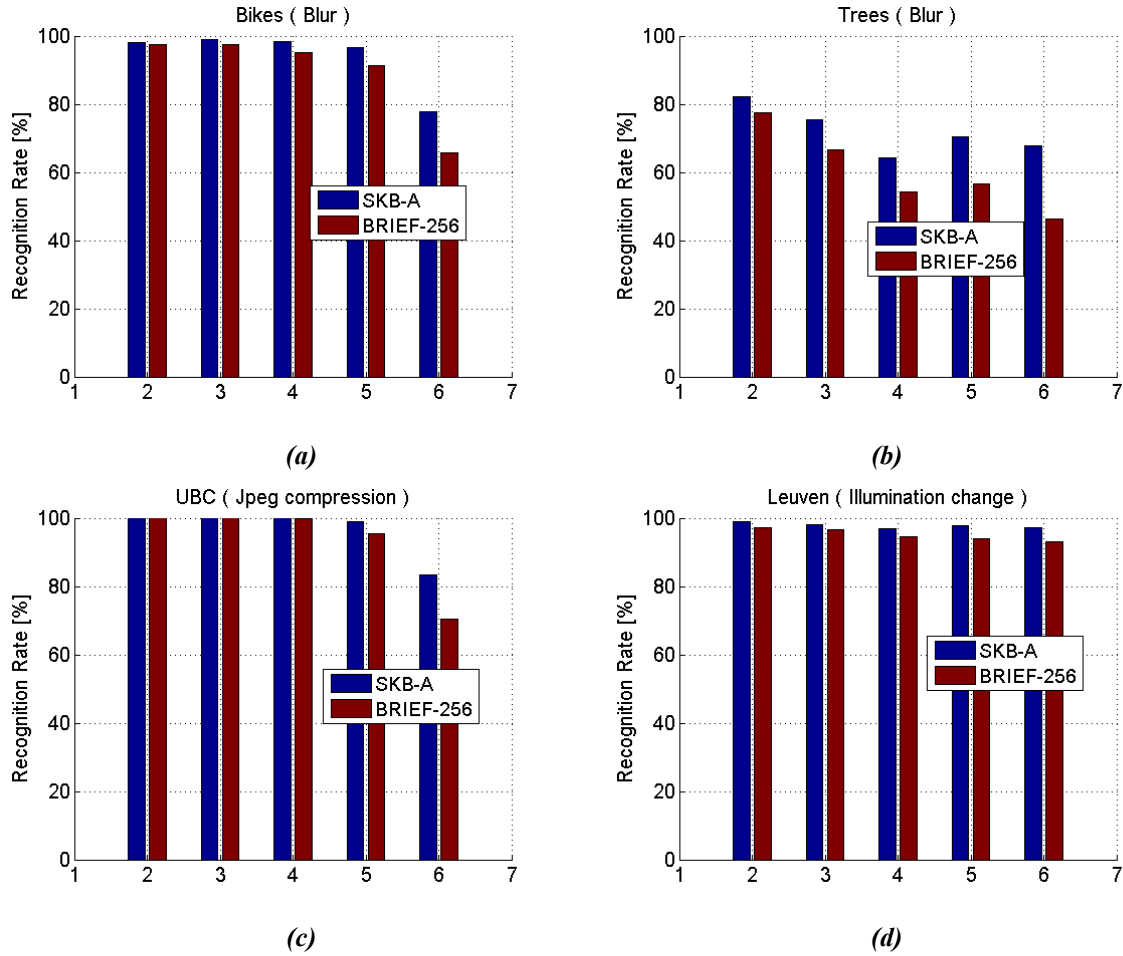


Figure 4.16. The SKB-A descriptor is compared to the BRIEF-256 descriptor using four image series. Each of the image series consists of 6 images, resulting in 5 pairs, i.e. image 1 is matched against the images 2 to 6. SKB-A shows a higher recall rate for all images.

4.6.3 Application to Real-Time Stereo Matching

In a third test, the SKB descriptor was used for a real-time stereo matching application. The video resolution was 960x540 pixels, the frame-rate was 25 Hz, i.e. 40 ms were available for the complete detection, description, and matching process. The interest point detector used in the real-time implementation is a variant of SUSurE which has been proposed in [Ebrahimi09] and described in section 2.5.2.

A single screenshot from the video sequence is shown in Figure 4.17. Approximately 3000 interest points were described per frame and camera, resulting in 600 consistent matches in average. The matching was performed using a left-right consistency check. Afterwards, a RANSAC algorithm was used to estimate the fundamental matrix and to subsequently eliminate outliers according to the algorithm described in sub-section 3.2.5. Table 4.2 shows the time consumption of the real-time matching application running on an Intel Xeon X5680 CPU at 3.33 GHz. The CPU has 6 cores plus 6 cores via hyper threading. The matching was performed using all CPUs, while the interest point detection and description was performed on 2 cores (one per camera). The interest point detection was

performed on a subsampled image with 480x270 pixels, while the pixel position interpolation according to the algorithms described in section 2.5.2.5 was conducted using the full resolution images with 960x540 pixels.

Table 4.2. Runtime of the different image processing steps. At 25 Hz, an image pair needs to be processed within 40 ms.

<i>Processing Step</i>	<i>Runtime in ms</i>
Integral Image	2
Binarized Laplacian of Gaussian	12
SKB Descriptors	3
Matching	12
RANSAC	8
Total	37

Table 4.2 shows that the real-time requirement was met within the stereo matching application. The description process and the matching needed 12 ms each. The description could be speeded up when using more than 2 CPUs for the process. However, even in this configuration, the stereo matching application ran fast enough for 25 fps.



Figure 4.17. Screenshot of the real-time stereo matching application. One can see that the scene contains objects at different distances to the stereo camera. The matched feature points are coloured according to their horizontal disparity, brown for near objects, blue for far objects. Apparently, no outliers are visible. The original resolution used for the matching was 960x540 pixels.

4.7 Conclusion

A new feature descriptor designed for real-time multi-camera applications which shows strong results at low outlier rates has been presented. The descriptor design is optimized for fast and robust correspondence search in multi-camera configurations. Its suitability in terms of robustness and speed were demonstrated in a comparison with the BRIEF descriptor as well as the SIFT and GLOH descriptor. Moreover the SKB was successfully used within a real-time stereo matching environment, underlying its suitability for real-time image processing applications.

The descriptor SKB was first proposed in [Zilly11c] and has since then attracted the interest of the research community. In that context, it was used by Stefanoski et al. [Stefanoski13] as part of the framework of the Image Domain Warping algorithm. As indicated in [Stefanoski13], the approach was submitted to MPEG resulting in one of the four best proposals in the multi-view-autostereoscopic display test scenario. Furthermore, the SKB was implemented as ASIC core using a SANDSTORM chip by Schaffner et al. in [Schaffner13]. The implementation has a complexity of 254 kGE and runs according to [Schaffner13] at 100 MHz being able to process 25000 Interest Points at 720p resolution in real-time.

In the context of this thesis, the SKB plays an important role for the stereoscopic 3D assistance system described in chapter 5. The feature detector based on the SKB is thereby used to analyze the scene geometry and to create a disparity histogram as will be shown in the subsequent chapter.

5 Assisted 3D Production

5.1 Introductory Remarks

The production of high quality stereoscopic 3D content remains a challenging task. As discussed in the introductory section 1.1.1, it is of great relevance for the development of the 3D-TV market to reduce the overall costs of a 3D production. In this context, state-of-the-art 3D production workflows have been described in section 1.2. Thereby, the quality of the produced 3D material needs to be kept high or even increased because improperly produced 3D can cause headaches and visual fatigue [IJsselsteijn00]. The geometry of 3D reproduction which is the mathematical link between the mechanisms of the human visual system on one hand and 3D production rules on the other hand was described in section 2.1. Basic insights about the human visual system and 3D perception which are related to visual discomfort have been described in section 2.2. The resulting 3D production rules were described in section 2.3. Given the introduction from chapter 1 and the theoretical background from chapter 2, the question is raised, what tools are needed for a stereographer to efficiently produce stereoscopic 3D content while taking important production rules into account.

Against this background, in this chapter, a technical solution for a more efficient production workflow using stereoscopic 3D cameras is described¹⁷. The basic idea is an assistance system which supports the stereographer on set and during post-production. In this context, the scene is analyzed in near real-time by the assistance system using input images from the left and the right camera. The underlying techniques used to accomplish this goal are the feature detector SKB from chapter 4 and the method for estimating the linearized fundamental matrix from chapter 3. They are applied to estimate the relative camera poses on one hand, and to analyze the scene depth structure on the other hand. The latter two analysis results can then be used to calibrate the stereo rig and to choose a proper convergence plane and inter-axial distance or to calculate temporally consistent rectifying homographies, e.g. for 3D live transmissions and preview purposes. Thereby, no expert knowledge in computer vision or epipolar geometry shall be required to use the assistance system. Consequently, an intuitive graphical user interface allows the stereographer to interact with the assistance system and to handle a set of comfort functions which facilitate the stereo 3D production workflow.

The remainder of this chapter is organized as follows: Related work is discussed in section 5.2 before an overview of the technical components of the assistance system will be given in section 5.3. In sections 5.4 and 5.5, the underlying algorithms for the time consistent analysis of the epipolar geometry and the depth structure of the scene are presented. In section 5.6, the implications of the comfort functions for the 3D production workflow are described and compared to the legacy stereo 3D production workflow. Finally, in section 5.7, the chapter will be concluded.

¹⁷ Parts of the content in this chapter have been previously published in [Zilly09], [Kauff10] and [Zilly10b].

5.2 Related Work

The first commercially available stereoscopic assistance system was the SIP2100 presented by 3ality digital [3ality] at IBC 2008 in Amsterdam. It consisted of a dedicated hardware system in 19" rack mount device. Being a commercial product, the available information regarding the functionality was limited to a specifications brochure. Seven months later at NAB show 2009 in Las Vegas, a first version of the stereoscopic analyzer (STAN) was presented which is described in this chapter. To the best of the author's knowledge the first mentioning in a scientific publication of a stereoscopic assistance system was in [Zilly09]. The short paper describes the concept of the assistance system described in this chapter. Publications in 2010 followed in a national technical journal [Kauff10] and in the proceedings of the International Conference on Image Processing [Zilly10b] describing a PC-based assistance system which is the reference system described in this chapter. Since then, several commercial assistance systems have been released [Binocle, Sony, Stereolabs]. Moreover, new assistance systems were proposed in the literature which refer to or discuss the presented assistance system. A computational stereo camera system closing the loop from analysis to motorized mechanical adjustment was proposed by Heinzle et al. [Heinzle11]. A system designed for interactive stereoscopic applications such as computer games was proposed by Oskam et al. [Oskam11]. An FPGA-based assistance system allowing also for the generation of disparity maps was proposed by Greisen et al. [Greisen11]. An approach for controlling the disparity in an interactive environment was proposed by Celikcan et al. [Celikcan13] where the authors also perform a user-study to identify comfortable parameter sets for convergence and inter-axial distance of the virtual camera system.

5.3 Overview of the Assisted 3D Production Setup

In this section, the main components of the proposed PC-based assistance system for stereoscopic 3D productions are described. Figure 5.1 gives an overview of the basic functionality of the assistance system. At the input the two camera streams are ingested into the system, e.g. via HD-SDI, Ethernet, or similar interfaces. The two camera signals are analyzed by an image processing algorithm and stereoscopic parameters of the analyzed scene are retrieved. The results of the analysis are visualized on the 3D viewfinder, a GUI which represents the interface for the interaction with the stereographer.

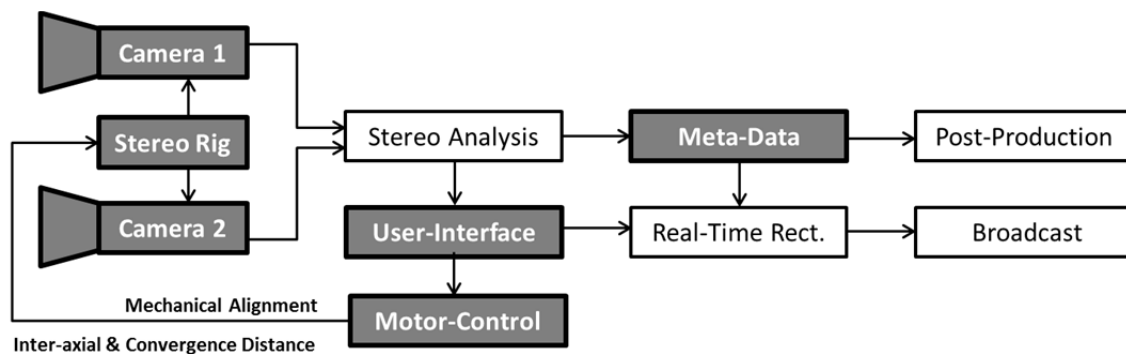


Figure 5.1. Proposed calibration and production workflow using the assistance system.

Main analysis results are the current depth volume, the position of the depth bracket, i.e. the convergence or zero-parallax plane, and specific geometrical parameters required for a mechanical alignment of the stereo rig such as roll and tilt angles or focal lengths of the two cameras. These parameters can be used to improve the alignment of the stereo rig. It can be done manually by a camera assistant, or automatically in case of a motorized rig and lenses. Furthermore, images from the adjusted stereo cameras are captured and analyzed permanently to close the loop of calibration and to monitor the accuracy of the mechanical alignment over time.

In addition, meta-data are generated to support correction of remaining geometric distortions by using image rectification. The correction can be done offline in post-production or in real-time for 3D live applications.

5.3.1 Stereo Rig with Mechanical Alignment Ability

When capturing stereoscopic content natively, i.e. using two cameras, a mechanical stereo rig is required as described in section 1.2.1. It is assumed that the rig is equipped with two cameras and suitable lenses, e.g. zoom lenses allowing the change of the focal length while shooting. It is assumed that the focal distance of the lenses can be changed which usually affects the focal length of the lenses as well. Ideally, the stereo rig itself offers different degrees of freedom through calibration plates to perform a mechanical alignment of the cameras. The calibration plates might even have a support for a motorized adjustment.

5.3.2 Frame Grabber

The Stereo Analysis is intended to run on a PC system, hence, the image data needs first to be captured. HD-SDI is a common standard for broadcast cameras and on-set monitors. It is specified by the SMPTE standardization body for different resolutions and frame-rates (e.g. SMPTE292M for 1080p25). Different PC extension cards exist which are able to capture the input from two HD-SDI cameras, e.g. DVS Atomix Board or BlackMagic 3D or from AJA Kona 3D. The footage intended for post-production generated using high quality cinema cameras is usually not captured using HD-SDI but for instance using inbuilt SxS cards. However, for the pure stereo analysis, the quality of the HD-SDI stream is sufficient. Alternatively, the data can be streamed into the PC system via Ethernet or similar interfaces.

5.3.3 Stereo Analysis Engine

The stereo analysis engine analyses the stereo images delivered by the frame grabber and generates meta-data, e.g. for a temporally consistent camera pose-estimation and depth volume analysis. The meta-data is then visualized using the GUI, stored for later post-production and sent to the real-time rectification unit. The latter needs to run at full frame rate while the stereo analysis might run at a lower frame rate, e.g. 5 frames per second while possibly being affected by considerable jitter. As consequence, a mechanism is required which performs a temporal filtering to ensure a temporally

consistent operation of the real-time rectification unit. The algorithms performed by the stereo analysis engine are presented in sections 5.4 and 5.5.

5.3.4 Real-Time Rectification Engine

The purpose of the real-time rectification engine is to eliminate vertical disparities at full frame rate. The rectification process needs to be temporally consistent in order to avoid jerky stereoscopic videos. The rectification engine might be implemented using an FPGA allowing very low delays which is important for many broadcast applications. Thereby, the complexity of the algorithms itself shall be kept as low as possible. Rectifying homographies are only applied within the engine, not calculated by the engine. It has to be taken into account that the input from the stereo analysis engine can jitter, thus rectifying homographies need to be interpolated componentwise while homographies for geometry correction and HIT (horizontal image translation) correction need to be combined efficiently.

5.3.5 Post-Production Unit

The geometric analysis and elimination of vertical disparities can also be performed during post-production. Thereby, the correction can be done using meta-data which has been generated and recorded during the shooting or using a stereo analysis which is also performed in post-production. Not all stereoscopic parameters can be changed within post-production. For instance, the analysis of the depth volume can be performed in post-production but the inter-axial distance cannot be changed anymore, unless involving complex algorithms such as disparity estimation and depth-image-based-rendering which might introduce new artifacts. The convergence plane in contrast can be changed with low effort in post-production. The same holds true for the elimination of vertical disparities.

5.3.6 Motor-Control Unit

If the stereo rig allows for the motorized adjustment of one or more degrees of freedom, a dedicated motor-control unit is used to interface between the results from the stereo analysis engine and the motors itself which might be affected by latencies, hysteresis or similar effects. Thereby the communication might require synchronous or asynchronous operation.

5.3.7 Graphical User Interface

A major role plays the 3D viewfinder which is implemented as a graphical user interface (GUI). It allows the visualization of important stereoscopic parameters as well as the presentation of the live images from the stereo cameras in different overlay modes. The 3D viewfinder can be displayed on a standard PC monitor or on a touch screen attached to the stereo rig. User input might be possible via touch screen, a conventional PC mouse, or dedicated hand controllers which can also steer some of the axes of a motorized stereo rig.

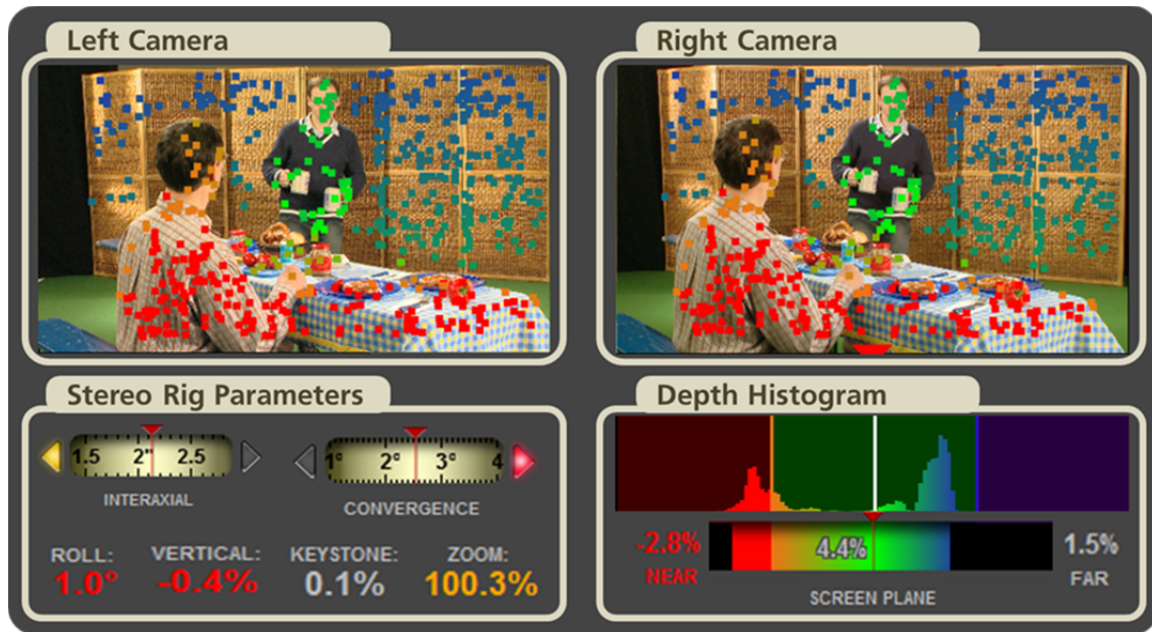


Figure 5.2. Schematic drawing of the graphical user interface (GUI). The left and right view, the feature points, stereo rig parameters and the depth histogram can be displayed.

As visualized in Figure 5.2, the results from the stereo analysis engine are displayed on the GUI which allows for a manual calibration of the stereo rig. Moreover, user input such as a wanted convergence plane, inter-axial distance, i.e. settings regarding the depth budget can be ingested via the GUI. The relationship between inter-axial distance and depth budget was described in sections 2.1 to 2.3. Different operation modes such as manual or automatic steering of the convergence distance and inter-axial distance can be selected via the GUI. Additional visualization modes such as the anaglyph representation format can be used for legacy 3D calibration.

5.4 Algorithms for the Geometry Analysis

In this section, the algorithms for a robust and time consistent analysis of the epipolar geometry are described.

5.4.1 Robust Feature Detection

The core element of the assistance system is a feature-point-based stereo analysis. The feature detector SKB from chapter 4 is used to find interest points and point correspondences between the two stereo images. Figure 5.3 illustrates this process. The green lines connect points found in the left and right image. In the ideal case these epipolar lines are horizontal and coincide with the image scanlines.



Figure 5.3. Point correspondences between same image regions are visualized as connecting lines. In the case of a rectified stereo pair, the connecting lines have only a horizontal component.

Obviously, any deviation from the ideal case results in the occurrence of unwanted vertical disparities. Using the theoretical framework of epipolar geometry, these deviations from the ideal state can be exploited to estimate the relative positions and orientations of the two cameras and their focal lengths. Following eqn. (3.10), this calculation might, for instance, discover the presence of an undesired roll of the camera with respect to the stereo baseline, or a deviation of the two focal lengths (cf. section 3.2.3 for details). Once estimated, the constraints of the epipolar geometry are used to identify robust matches. Outliers which do not fit into these constraints are discarded. This is done using a RANSAC computation of the fundamental matrix F which has been described in section 3.2.5.

Algorithm 5.1. FIFO buffered feature points with RANSAC filtered input

```

Input :      Handle to frame buffer FRAME_BUFFER_ID

Output :    List of feature points FP_LIST

struct FEATUREPOINT{age,u1,u2,v1,v2}
FP_LIST=create_empty_list (FEATUREPOINT)
NEW_FP=create_item (FEATUREPOINT)

do      Iterate until receiving stop signal
  for each FP in FP_LIST      Iterate over feature point list
    FP.age = FP.age+1;        Increase age of all FPs
    if FP.age > MAX_AGE
      remove FP from FP_LIST
    end
  end for
[imgL, imgR]=get_newest_images_from_framebuffer (FRAME_BUFFER_ID)
[u,v,u',v']=get_SKB_matches ([imgL, imgR]) ;      Array with size  $n \times 4$ , where  $n$  is number of feature points
matches_SKB = [u,v,u',v'];
matches_filtered = get_RANSAC_filtered_matches(matches_SKB) ;
for i=1 to size(matches_filtered)      Iterate over feature point list
  NEW_FP.u1 = matches_filtered[i][1];      // element u
  NEW_FP.v1 = matches_filtered[i][2];      // element v
  NEW_FP.u2 = matches_filtered[i][3];      // element u'
  NEW_FP.v2 = matches_filtered[i][4];      // element v'
  NEW_FP.age = 0;
  NEW_FP.image_width=get_width (imgL) ; // required for later normalization
  FP_LIST.insert (NEW_FP);
end for
until STOP_SIGNAL == TRUE

```

To enhance the robustness of later geometry analysis algorithms, feature point matches from multiple image pairs are gathered in a ring buffed list which is constantly updated. It is assumed that the epipolar geometry might change from frame to frame in a limited range. As described in Algorithm

5.1, the RANSAC filtering is therefore performed on each analyzed image pair, but the entire list of filtered feature points from multiple image pairs is used for analyzing the scene geometry.

5.4.2 Temporally Filtered Pose Estimation

The estimation of the relative pose of the two stereo cameras is the core component of the mechanical alignment assistance. Thereby, temporally stable results are easier to interpret by the user than jittering values which refer to possible alignment errors. In this sub-section, the three main sources for possible jitter and suitable strategies for a temporal filtering are described.

The camera pose estimation is done using feature points which might inhibit outliers which can be scene dependent and thus vary in a dynamic scene. Image regions with repetitive patterns for instance can increase the amount of outliers. In general the outliers are less stable in terms of matching score than inliers and hence show a stronger variance over time. The amount of outliers versus inliers is denominated as precision and is an important quality measure for a feature detector and descriptor. Details regarding the precision and recall rate of different feature descriptors are given in section 4.6.

Another source for unstable feature points which occurs even in the case of a static camera is image noise. It can affect the number and position of interest points detected using blob detectors as described in section 2.5 and chapter 4. In fact, image noise can influence the blobness of a pixel region pushing it above or below a specific threshold from frame to frame. Subsequently, the content of the keypoint descriptor might also vary due to pixel noise in the video sequence. Consequently, the respective matching score from putative matches between two stereo images might vary.

Yet another source for a temporally unstable set of feature point matches is the RANSAC filter which is used to eliminate those feature points from a set which do not fulfill the epipolar constraint as described in section 3.2.5. A predefined number of iterations is performed during the filtering process while a small set of randomly chosen feature points is chosen for each iteration of the RANSAC. It is the involved random number generator which yields to inherently temporally unstable results.

Given the three sources for temporally inconsistent feature point lists, the question is raised, how a temporally consistent pose estimation can be implemented. In fact, a set of temporal filtering steps can be applied. The first step is to use feature point matches from several analyzed frames for the pose estimation, i.e. to use a running average of feature points for the scene geometry analysis. A suitable algorithm involving a FIFO-buffer was described in the previous sub-section 5.4.1.

Given the set of feature points, the fundamental matrix describing the relative position of the two cameras is estimated using the approach described in section 3.2.3 i.e. based on the linearized fundamental matrix defined in eq. (3.9) and taking into account the epipolar constraints given by a stereoscopic setup with two nearly rectified cameras. A result vector \mathbf{x} according to eqn. (3.15) is extracted as a solution of the set of linear equations from eqn. (3.14). Assuming that the focal length f

is known, the relative angles and camera position can directly be extracted. After eliminating f from eqn. (3.15) the components are regrouped. Subsequently, the following normalized result vector is calculated which contains the required intrinsic and extrinsic parameters:

$$\mathbf{x}_{norm} = (\alpha_x, \alpha_y, \alpha_z, \alpha_f, \hat{c}_y, \hat{c}_z)^T. \quad (5.1)$$

Please note that the meaning of the different variables is explained in section 3.2 and eqn. (3.11) in particular. By $\alpha_x, \alpha_y, \alpha_z$ the angles between the orientation of the right camera and the x -, y - and z -axis respectively are described. They vanish in the case of rectified cameras, e.g. the residual angles denote a deviation of the right camera from its optimal orientation. The term α_f denotes the relative difference between the two focal lengths, e.g. it vanishes when the focal lengths from the left and the right camera are identical. With \hat{c}_y and \hat{c}_z translational errors of the camera's centers are denoted.

The result vector \mathbf{x}_{new} can be smoothed by linearly combining a result vector from the last geometry estimation \mathbf{x}_{old} with the newest result vector \mathbf{x}_{norm} from eqn. (5.1). The ratio between the previous and the current geometry estimates can be controlled using a constant α as shown in the following equation:

$$\mathbf{x}_{new} = \alpha \cdot \mathbf{x}_{old} + (1 - \alpha) \cdot \mathbf{x}_{norm}. \quad (5.2)$$

The optimal choice of the constant α depends on the amount of noise in the process of geometry estimation, which is generated due to outliers within the feature detection process, image noise, and stochastic processes performed during the RANSAC fitting. A larger α will lead to a stronger temporal smoothing, but also to a stronger delay until the stabilized value has been calculated.

5.4.3 Temporally Consistent Stereo Image Rectification

Remaining vertical disparities or other geometrical distortions which are undesired for any 3D production in general and for 3D live broadcast in particular can be eliminated using a stereo image rectification process. The effect of image rectification is shown in Figure 3.2. In case of a live broadcasting this is a probate method to ensure a high quality of the stereoscopic live stream. Even after a careful calibration, it might be necessary to apply an image rectification, when for instance internal or external camera parameters have been changed. The usage of zoom lenses is such a scenario. The principal point of zoom lenses moves when changing the focal lengths [Fraser06, Fobker11]. In stereo applications this effect results in vertical disparities changing dynamically while zooming. Furthermore, some small-size stereo rigs do not offer the possibility for a mechanical alignment of all degrees of freedom. In that case, image rectification is the only way to avoid image distortions.

The algorithm for the derivation of the rectifying homographies for a static scene was described in section 3.2.7. In order to extend this algorithms towards a real-time capable stereo rectification for dynamic scenes, a few additional requirements need to be considered.

5.4.3.1 Requirements

Most importantly, the homographies which are applied to consecutive stereo frames should not vary too strongly, because this could lead to strong jitter artifacts and visual discomfort [Templin14]. This has to be ensured even if the update rate of the stereo analysis varies and is lower than the video frame rate. It has to be taken into account that the stereo analysis module (see sub-section 5.3.3) and the rectification engine (see sub-section 5.3.4) are two separate modules which might not run synchronously. The latter shall avoid complex algorithms to allow an easy implementation in hardware and thus allowing very low latency processing which is important for live applications. Finally, the adaptation of the convergence plane shall be performed in a single step along with the rectification process. The corresponding horizontal image translation (HIT), introduced in section 2.1.3, might be chosen manually or computed according to dedicated algorithms which will be described in sub-section 5.5.2. In any case, the update rate for the HIT might differ from the one of the stereo geometry analysis.

5.4.3.2 Approach

The rectifying image transforms, the homographies, are derived from the temporally filtered vector \vec{x}_{new} which was calculated according to eqn. (5.2) by the stereo analysis engine. The derivation of homographies given a set of geometric parameters was described in section 3.2.7. The resulting homographies are then sent to the real-time rectification engine. The adaptation of the convergence plane can also be expressed using a homography according to eqn. (5.8) as will be shown in section 5.5.2. The respective homographies can be calculated according to an algorithm which will be explained in the subsequent section.

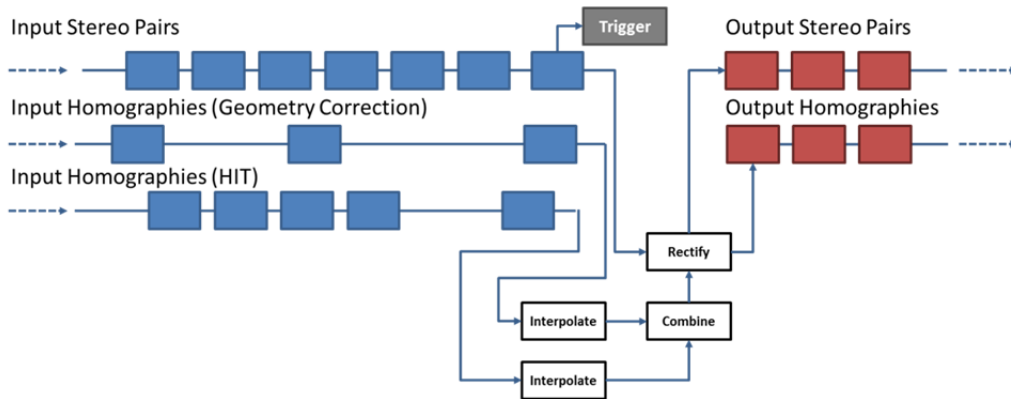


Figure 5.4. Overview of the temporally consistent combination and interpolation of homographies.

It is important for the real-time rectification engine that two sets of homographies with independent update rate are ingested along with a continuous stream of stereo image pairs. This is visualized by the

three inputs lines with blue data packets in Figure 5.4. In the simplest case, one could apply the most recent homography pairs which were received to perform the image rectification and adaptation of the convergence plane. This however, would result in a non-continuous rectification with strong jumps in the convergence plane which might impair the visual comfort [Templin14]. To avoid this, the two input homographies pairs need now to be interpolated for each input stereo image pair. Each blue input homography from Figure 5.4 can be seen as a new target homography \mathbf{H}_{target} , one for the rectification and one for the convergence plane adaptation. This target homography is now compared with the current homography $\mathbf{H}_{current}(t-1)$ applied to the last frame to calculate a new current homography $\mathbf{H}_{current}(t)$ according to eqn. (5.3).

$$\mathbf{H}_{current}(t) = \mathbf{H}_{current}(t-1) + \alpha \cdot [\mathbf{H}_{target} - \mathbf{H}_{current}(t-1)]. \quad (5.3)$$

The interpolation of the homographies is done component-wise, e.g. for each of the nine elements of the 3x3 matrices. The parameter α controls the temporal smoothing of the homographies and needs to be in the following range: $\alpha \in]0, 1]$. Please note that this approach is not suitable for homographies in the general case as trigonometric functions are involved. However in the case of rectifying homographies \mathbf{H}_{rect} calculated for cameras which are near the rectified state, interpolating homographies componentwise, behaves similarly to the addition and subtraction of small angles which can be linearized following the justification given in section 3.2.2. The homographies for the horizontal image translation \mathbf{H}_{HIT} as defined in eqn. (5.8) can be interpolated component-wise without loss of generality.

After the interpolation, the homographies for the geometric correction and the convergence plane adjustment can be combined in the following way using a simple matrix multiplication:

$$\mathbf{H} = \mathbf{H}_{HIT} \cdot \mathbf{H}_{rect}. \quad (5.4)$$

Finally, the homographies which are now interpolated and combined can be applied to the stereo image pairs as illustrated in Figure 5.4. The real-time rectification engine outputs temporally consistent stereo image pairs with the respective homographies, illustrated as red data packets.

5.5 Algorithms for the Depth Bracket Analysis

The results of the feature-based stereo analysis can also be used to derive the near and the far clipping plane of the scene. These are important stereoscopic parameters which are related to the depth volume and the convergence plane (cf. also section 2.3). The former implies the width of the depth bracket while the latter defines the position of the depth bracket in 3D space [Mendiburu12].

5.5.1 Calculation of a Disparity Histogram

Within the stereo analysis engine from the previous section, a list of feature point correspondences is generated along with rectifying homographies. These feature points shall now be used to calculate the width and position of the depth bracket, i.e. the near- and far-clipping plane. For each pair of feature points, $\mathbf{m}_i = (u_i, v_i, 1)^T$ and $\mathbf{m}'_i = (u'_i, v'_i, 1)^T$ the respective horizontal disparity $u'_i - u_i$ can be calculated.

The simplest approach, given a set of n feature point pairs, to calculate the minimum disparity $disp_{min}$, the maximum disparity $disp_{max}$, and the disparity range $disp_{range}$ is the following:

$$\begin{aligned} disp_{min} &= \min\{u'_1 - u_1, u'_2 - u_2, u'_3 - u_3, \dots, u'_n - u_n\}, \\ disp_{max} &= \max\{u'_1 - u_1, u'_2 - u_2, u'_3 - u_3, \dots, u'_n - u_n\}, \\ disp_{range} &= disp_{max} - disp_{min}. \end{aligned} \tag{5.5}$$

However, this approach has a few disadvantages which will be discussed shortly. First, it is very susceptible to outliers, e.g. a single mismatch in the feature points could have a direct and strong impact on the calculated minimum or maximum disparity. Moreover, the localization precision of feature points is finite and tends to vary with pixel noise in the stereo images used for the feature point detection. As consequence, feature points pairs which, for instance, should have the same disparity $d_i = u'_i - u_i$ give raise to a Gaussian distribution of disparity values where the standard deviation σ depends on the quality of the feature detector and the input images. The minimum disparity $disp_{min}$ tends therefore to be underestimated while the maximum disparity $disp_{max}$ tends to be overestimated along with the disparity range $disp_{range}$. Finally, geometric distortions in the stereo images induced by mechanical misalignment do not only induce vertical disparities but also disturb the horizontal disparities and the calculated disparity range.

5.5.1.1 Approach

The challenges described above are taken into account by modifying the naïve approach from the previous sub-section in the following way: First, the feature point pairs from the robust feature detection are rectified using the rectifying homographies. The homographies to be used were estimated in the context of the temporally consistent stereo image rectification from sub-section 5.4.3. Subsequently, the horizontal disparities calculated using the rectified feature points are gathered in a disparity histogram. To eliminate outliers, the lowest and highest percentiles are discarded from the histogram, a value which can be parameterized using the parameter $perc_{outlier}$ in the algorithm described below.

Algorithm 5.2. Rectification-aware Disparity Histogram Calculation

Input : List of feature points: FP_LIST
rectifying homographies: H, H'
disparity ranges: $bin_disp_min, bin_disp_max$
outlier percentiles: $perc_{outlier}$
feature point noise: σ_{fp}
temporal smoothness: σ_{age}

Output : $disp_{min}, disp_{max}, disp_{range}$
Disp_histogram

```

bin_count = 101; minimum bin_count = 2
bin_stepsize = (bin_disp_max - bin_disp_min) / (bin_count - 1);
disp_histogram = create_empty_histogram(bin_count);
weight_sum = 0;
// create weighted histogram
for each FP in FP_LIST Iterate over feature point list
    fp_weight = exp(-FP.age /  $\sigma_{age}$ ); Histogram weight depends on age
    [u, v, s] = [FP.u1, FP.v1, 1.0];
    [u', v', s'] = [FP.u2, FP.v2, 1.0];
    [ $\hat{u}, \hat{v}, \hat{s}$ ] = apply_homography([u, v, s], H);
    [ $\hat{u}', \hat{v}', \hat{s}'$ ] = apply_homography([u', v', s'], H');
    disp = ( $\hat{u}' / \hat{s}' - \hat{u} / \hat{s}$ ) / FP.image_width; // normalize disparities w.r.t. to image width
    Bin_index = floor((disp - bin_disp_min) / bin_stepsize);
    if (bin_index > 1) & (bin_index <= bin_count)
        disp_histogram[bin_index] += fp_weight;
        weight_sum += fp_weight;
    end
end for
// retrieve bin corresponding to n-th percentile from histogram and subtract margin
sum_low = 0; sum_high = 0;
for i = 1:bin_count
    sum_low += disp_histogram[i];
    if sum_low / weight_sum >=  $perc_{outlier}$ 
         $disp_{min} = (i - 1) * bin\_stepsize + bin\_disp\_min + \sigma_{fp}$ ;
        break;
    end
end
for i = 1:bin_count
    sum_high += disp_histogram[bin_count - i + 1];
    if sum_high / weight_sum >=  $perc_{outlier}$ 
         $disp_{max} = (1 - i) * bin\_stepsize + bin\_disp\_max - \sigma_{fp}$ ;
        break;
    end
end
if  $disp_{max} < disp_{min}$  // can happen when overestimating  $\sigma_{fp}$ 
     $disp_{min} = (disp_{max} + disp_{min}) / 2$ ;
     $disp_{max} = disp_{min}$ ;
end
disp_range =  $disp_{max} - disp_{min}$ ;

```

The statistical weight of the feature points depends on their position within the FIFO buffer (cf. Algorithm 5.1). The temporal smoothing can be influenced by the factor σ_{age} used in Algorithm 5.2.

A value of 2% means that 2% of the histogram are cut from the histogram and that the 3rd (98th) percentile is identified as minimum (maximum) disparity value. Finally, in order to account for noise in the pixel positions of the feature points, a value parameterized using σ_{fp} is subtracted from the minimum and maximum disparities.

As output of the algorithms, the minimum disparity value $disp_{min}$ which corresponds to the near clipping plane, and the maximum disparity value $disp_{max}$ corresponding to the far clipping plane along with the disparity range $disp_{range}$ are calculated. The disparities have been normalized with respect to the image width to get resolution-independent values.

5.5.2 Automatic Adjustment of the Horizontal Image Translation (HIT)

The perceived convergence plane on a stereoscopic 3D device can be changed by applying a horizontal image translation (HIT) to the stereo images or videos regardless of the camera setup used for shooting the scene, i.e. using parallel or convergent optical axes (cf. section 2.1.3). The aim of the automatic adjustment of the HIT is to keep the disparities occurring in the scene within the comfortable viewing range (cf. section 2.3), a task which is comparable to the one of the convergence puller. The optimal HIT is therefore scene dependent, dynamic, and depending on the viewing conditions such as display size. The parameter which interacts between these settings and the proposed HIT adjustment algorithm is the depth budget (cf. Table 2.1), and more precisely the minimum (negative) parallax $\tilde{\mathcal{P}}_{min}$ and the maximum (positive) parallax $\tilde{\mathcal{P}}_{max}$ expressed as percentage of the screen width.

Now, assuming that a stereo image pair covers the full surface of the screen, it can be concluded that a comfortable viewing experience is given as long as the following inequalities hold true:

$$\begin{aligned} disp_{min} &\geq \tilde{\mathcal{P}}_{min} \\ disp_{max} &\leq \tilde{\mathcal{P}}_{max} \end{aligned} \tag{5.6}$$

Please note that the terms $\tilde{\mathcal{P}}_{min}$ and $\tilde{\mathcal{P}}_{max}$ refer to the available depth budget from section 2.3 while the terms $disp_{min}$ and $disp_{max}$ refer to the measured depth volume in the scene, e.g. the disparities as percentage of the sensor width.

Action is required if the measured disparities exceed the available depth budget. The idea is to add a horizontal image translation HIT which brings the measured disparities back into the comfortable viewing range. However, this is only possible as long as the measured disparity range $disp_{range} = disp_{max} - disp_{min}$ fits within the maximum and minimum allowed parallax values, e.g. as long as the following inequality holds true:

$$disp_{max} - disp_{min} \leq \tilde{\mathcal{P}}_{max} - \tilde{\mathcal{P}}_{min} \tag{5.7}$$

If the above condition is not met, a heuristic is required which decides whether it is better to exceed the allowed minimum parallax $\tilde{\mathcal{P}}_{min}$ or the maximum parallax $\tilde{\mathcal{P}}_{max}$, i.e. which of the rules from section 2.3.1 has priority. For an educated answer, the consequences for the 3D perception and the human visual system have to be taken into account as well as the viewing conditions. Exceeding the

minimum parallax $\tilde{\mathcal{P}}_{min}$ might result in an accommodation-convergence conflict (**Rule 2**) or in a window or framing violation (**Rule 1**). Exceeding the maximum parallax $\tilde{\mathcal{P}}_{max}$, however, might result in a configuration where the viewer's eyes need to diverge (**Rule 3**) which is a very stressful situation. Hence in that case priority should be given to the maximum parallax. The above mentioned rules are taken into account by the heuristic presented in Figure 5.5.

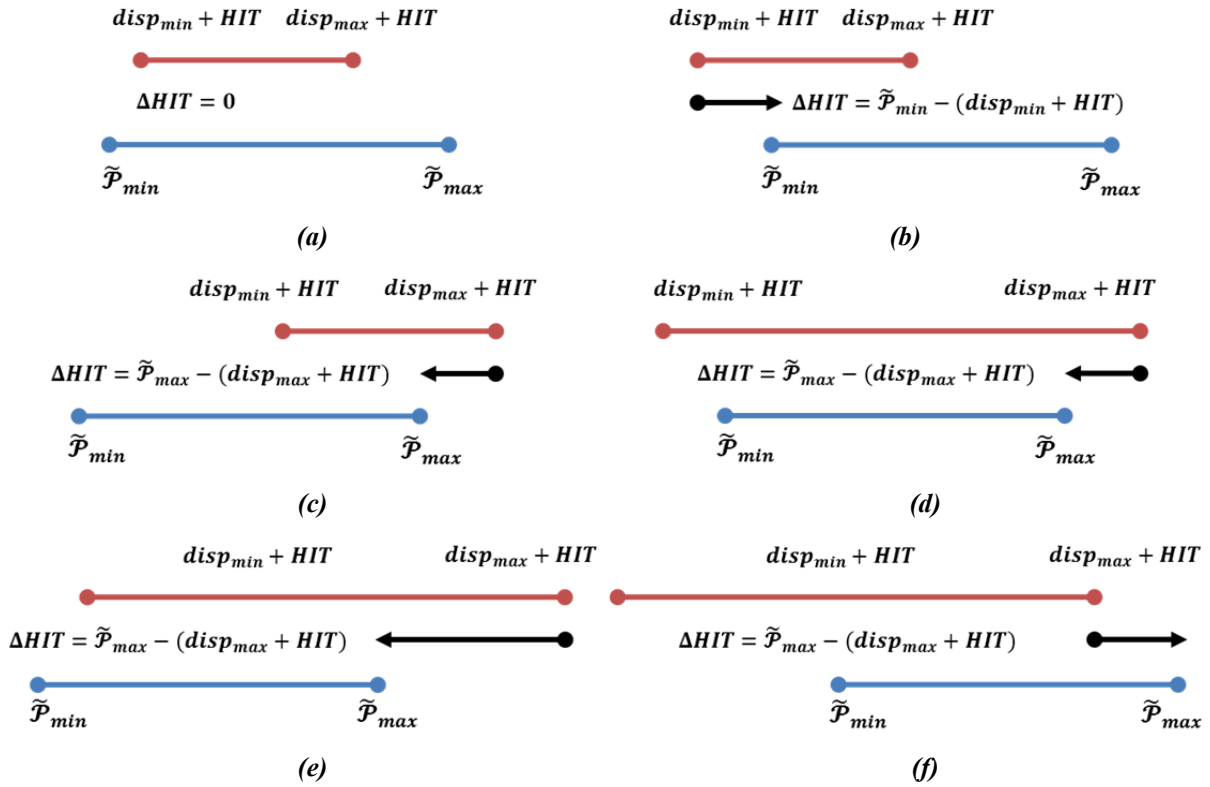


Figure 5.5. Proposed heuristic for the dynamic adjustment of the horizontal image translation (HIT).

Six different configurations are distinguished within the heuristic which are presented in the sub-images (a) to (f) of Figure 5.5. The range between the minimum and parallax values is indicated in blue, which refers to the available depth budget or comfortable viewing range. The range indicated in red refers to the measured disparity volume. Ideally, it should fit within the blue range which is the case in (a), where no adjustment of the HIT is necessary ($\Delta HIT = 0$). It could be useful to center the depth volume within the depth budget. However, this would mean that the convergence plane would be adjusted more often as necessary which would reduce the temporal consistency of the convergence plane. In the cases (b) and (c), an adjustment of the position of the so-called depth bracket using the *HIT* allows to move the depth volume into the comfortable viewing range by increasing (b) or decreasing (c) the *HIT* value in order to respect the minimum parallax value (b) or the maximum parallax value (c). The situation differs in the cases (d), (e) and (f) as the measured depth volume exceeds the allowed depth budget, i.e. the condition from eqn. (5.7) is not met. As explained above, the priority is now to make sure that the maximum parallax values $\tilde{\mathcal{P}}_{max}$ are not exceeded, giving priority to **Rule 3** from eqn. (2.5). The *HIT* is adjusted accordingly in the respective cases (d), (e), and

(f) from Figure 5.5. Please note that a reduction of the disparity range cannot be performed by applying a *HIT* but by decreasing the inter-axial distance, i.e. the stereo baseline \mathcal{B} as will be described in the next sub-section.

Once the *HIT* value has been evaluated or updated, corresponding homographies \mathbf{H}_{HIT} and \mathbf{H}'_{HIT} are generated for the left and the right camera according to eqn. (5.8):

$$\mathbf{H}_{HIT} = \begin{bmatrix} 1 & 0 & \frac{HIT}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{H}'_{HIT} = \begin{bmatrix} 1 & 0 & -\frac{HIT}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (5.8)$$

As can be seen, the horizontal image translation is equally distributed to the left and the right camera images.

Finally, the updated homographies \mathbf{H}_{HIT} and \mathbf{H}'_{HIT} are sent to the real-time rectification engine in order to be combined in a temporally consistent way with the rectifying homographies as described in section 5.3.4.

5.5.3 Automatic Derivation of the Inter-axial Distance

The depth histogram and the derived minimum and maximum disparity values $disp_{min}$ and $disp_{max}$ can be used to estimate the near and far clipping plane and to derive the current depth volume as described in the previous section. The depth volume $disp_{max} - disp_{min}$ is compared to the available depth budget $\tilde{\mathcal{P}}_{max} - \tilde{\mathcal{P}}_{min}$ and as a result an optimized inter-axial distance or stereo baseline¹⁸ $\mathcal{B}_{optimal}$ can be derived given the current inter-axial distance $\mathcal{B}_{current}$. In the case of rectified cameras, the depth volume is proportional to the inter-axial distance, i.e. by increasing the inter-axial distance by e.g. 10%, the depth volume is increased by the same amount. Consequently, when the ratio of current and optimal depth volume is known, the updated inter-axial distance can easily be derived according to the following equation [Zilly2009]:

$$\mathcal{B}_{optimal} = \mathcal{B}_{current} \cdot \frac{\tilde{\mathcal{P}}_{max} - \tilde{\mathcal{P}}_{min}}{disp_{max} - disp_{min}}. \quad (5.9)$$

In order to know in which direction the inter-axial has to be changed, the difference between the optimal and the current baseline can be derived as follows:

$$\Delta\mathcal{B} = \mathcal{B}_{current} \cdot \left(\frac{\tilde{\mathcal{P}}_{max} - \tilde{\mathcal{P}}_{min}}{disp_{max} - disp_{min}} - 1 \right). \quad (5.10)$$

¹⁸ In this section, the terms *baseline* and *inter-axial distance* are used interchangeably.

This heuristic can now be applied to modify the inter-axial distance as illustrated in Figure 5.6. If the measured depth volume $disp_{max} - disp_{min}$ is smaller than the available depth budget $\tilde{\mathcal{P}}_{max} - \tilde{\mathcal{P}}_{min}$, the stereo baseline or inter-axial distance needs to be increased, i.e. $\Delta\mathcal{B} > 0$ as indicated in Figure 5.6 (a). If the inverse is true, as shown in Figure 5.6 (b), the inter-axial distance needs to be decreased.

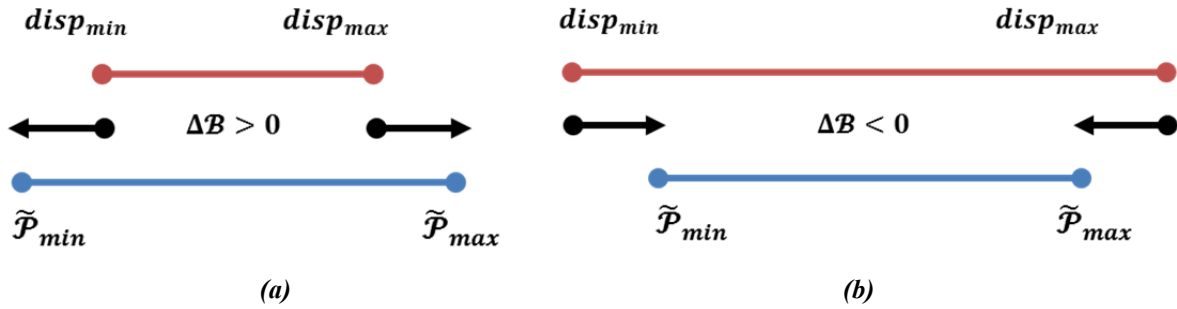


Figure 5.6. Modification of the inter-axial distance or stereo baseline \mathcal{B} according to the ratio of the measured depth volume and the available depth budget.

Although the target inter-axial distance can be calculated, depending on the speed of the motors and taking into account a certain latency of the whole system, it can be useful to modify the inter-axial distance only by a small amount, e.g. 0.5 mm. In this case, only the signum of the value $\Delta\mathcal{B}_{baseline}$ needs to be known and not the exact value of the current baseline. Moreover, in order to ensure a time-consistent inter-axial distance even in the case of dynamic scenes, the baseline should only be modified if the difference between current and optimal baseline exceeds a threshold $\theta_{baseline}$. Finally, a correspond command needs to be sent to the motor-control unit described in section 5.3.6.

5.6 Comparison with Legacy Workflows

The purpose of the set of algorithms presented in the previous section is to facilitate the work of the stereographer in general, and to give assistance to meet the stereoscopic production rules from section 2.3 in particular. Consequently, in this section the workflow improvements which can be realized compared to legacy production workflows will be described.

5.6.1 Stereo Rig Calibration and Rectification

5.6.1.1 Manual Stereo Calibration Workflow

Although different calibration workflows exist, a workflow involving the use of a checker-board test chart will be used as reference workflow. It is assumed that a live picture of the left and right camera is available which can be visualized as anaglyph composite image. Alternative visualization modes such as the difference of the two images or the overlay of the two images are also widely used by stereographers [Mendiburu2012].

When the two cameras are misaligned, this induces vertical and horizontal disparities which can be identified by the experienced stereographer. What adds to the difficulty of the task is that all effects overlay which makes it difficult to identify the different sources for the disparities. Consequently, an

iterative approach is favorable where first all axes are roughly aligned, before a refinement of all angles and parameters is conducted. It is assumed that the inter-axial distance axis and the convergence axis are roughly aligned in a first step before starting the alignment of the remaining angles.

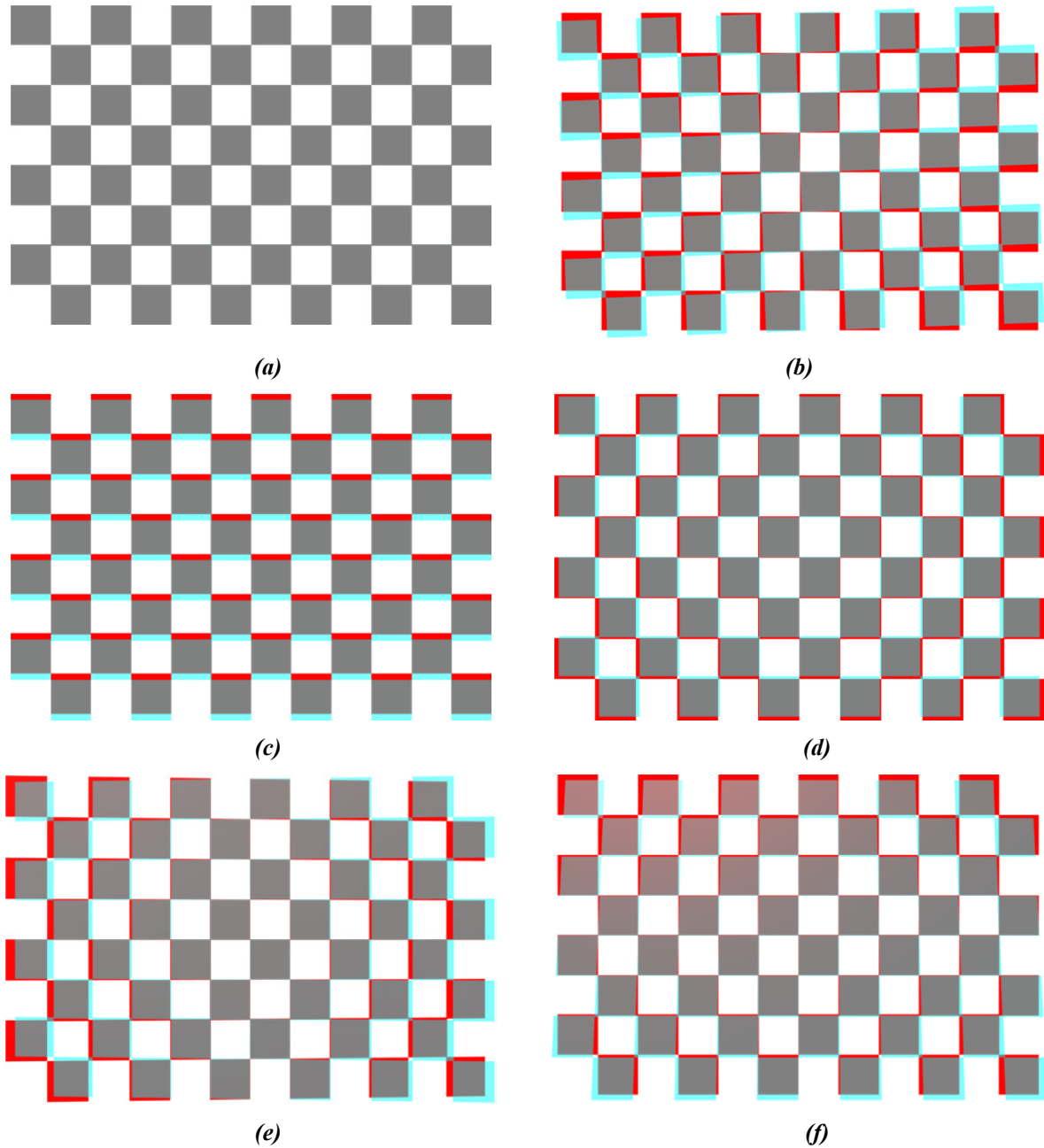


Figure 5.7. *Mechanical alignment of a stereo rig using a checkerboard test pattern and anaglyph visualization mode. (a) Ideal case where no differences are visible, position, orientation and intrinsic parameters match. (b) A roll error around the z-axis yields to this effect. (c) Vertical misalignment due to difference of the y-position of the principal point or mismatch of the tilt angles around the x-axis. (d) A difference of the focal lengths yield to this effect, or translation error in z-direction if a single depth plane is analysed. (e) Keystone-effect arising from non-parallel, i.e. converging or diverging optical axes yielding to vertical and horizontal disparities. (f) Keystone-effect in vertical direction induced by a mismatch in the tilt-angle between the two cameras.*

Different effects arising from mechanical misalignment of the stereo rigs are illustrated in Figure 5.7. It is assumed that the stereographer minimizes the different orientation errors step by step. Once a

minimum has been reached, the exact zero-position of the inter-axial distance and a possible translation error in y -direction can be calibrated. This requires an additional horizontal bar for the y -position and a vertical bar for the x -position which need to be placed in a different depth plane than the checker-board in order to generate parallax between the two objects.

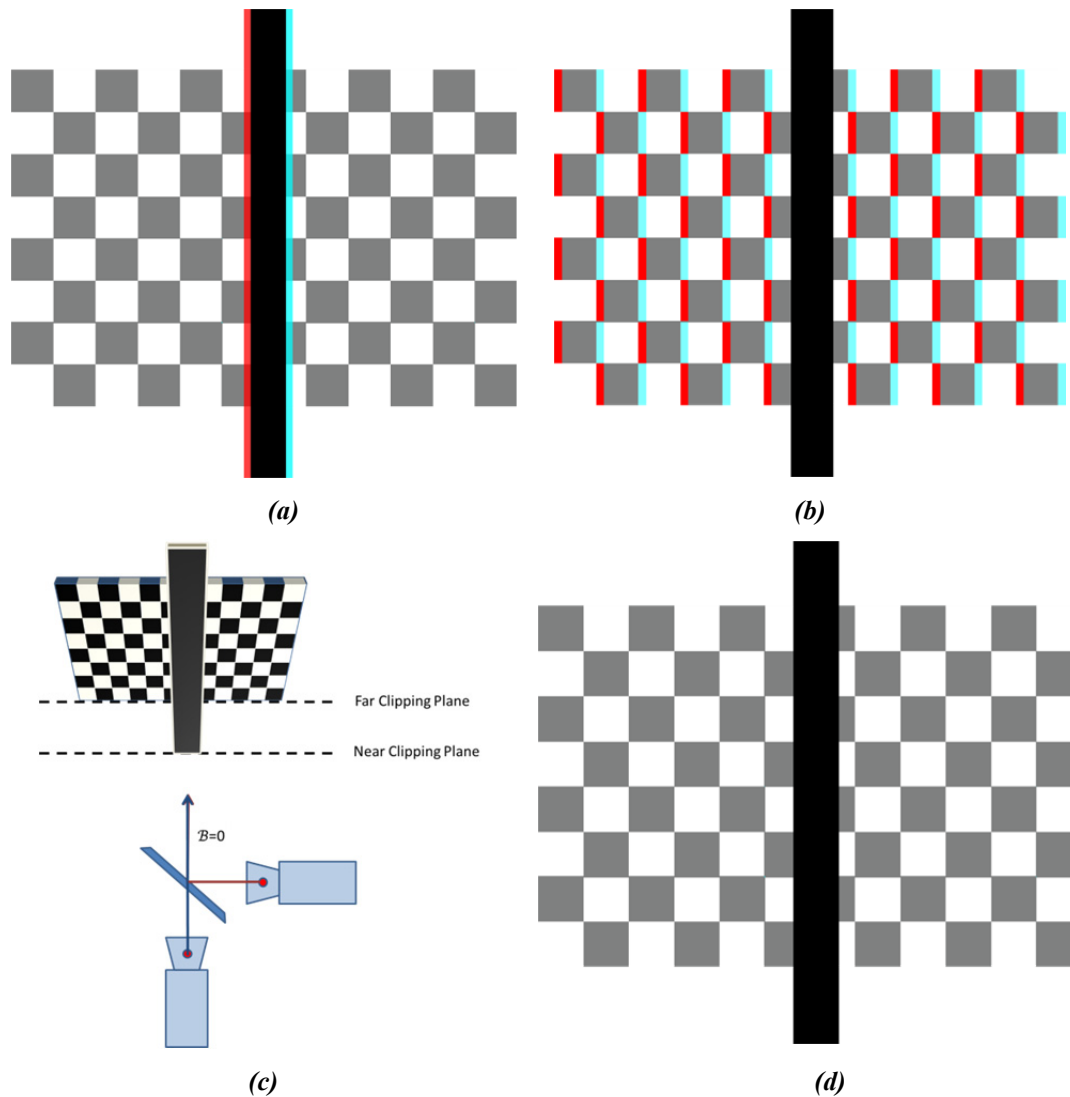


Figure 5.8. Setup for the calibration of the zero-parallax position for the inter-axial distance. (a,b) The inter-axial distance is not zero, hence horizontal disparities can be eliminated for the checkerboard in the background (a) or the black bar in the foreground (b), but not both at same time. (c) Illustration of the optical axes and the inter-axial distance B when using a beam-splitter rig. (d) The inter-axial distance is zero, hence horizontal disparities can be eliminated for the fore- and background at the same time.

The setup is illustrated in Figure 5.8. As long as there is a parallax between the left and the right camera, it is not possible to eliminate the horizontal disparities in the foreground and background at the same time as shown in Figure 5.8 (a) and (b). However, if the inter-axial distance B is zero as shown in Figure 5.8 (c), all horizontal and vertical disparities can be eliminated as shown in Figure 5.8 (d).

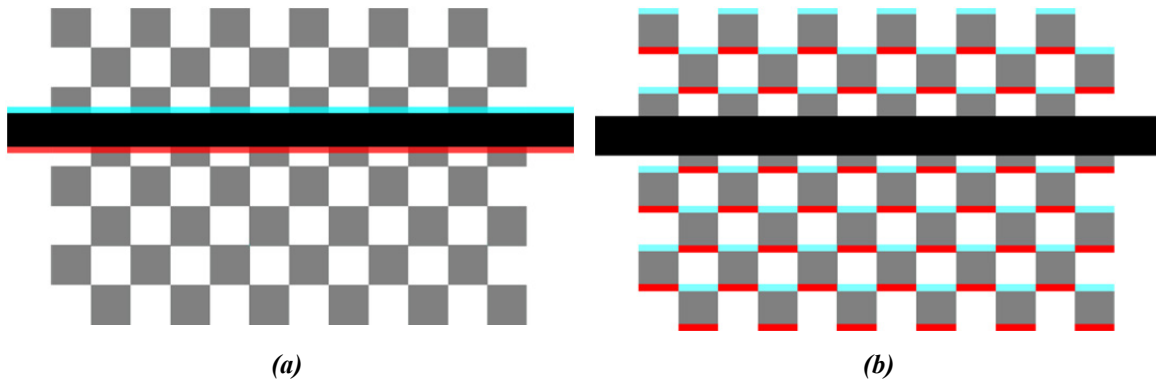


Figure 5.9. Effect of a vertical translation error, e.g. in y -direction. The vertical disparities can only be eliminated for the checkerboard in the background (a), or for the black bar in the foreground (a), but not for both objects simultaneously.

In order to calibrate the y -position of the two cameras and to minimize translational errors, a horizontal bar can be chosen as foreground object as shown in Figure 5.9. As long as translational errors are present, vertical disparities can be eliminated only for the background as shown in Figure 5.9 (a) or the foreground as shown in Figure 5.9 (b), but not both at the same time.

Finally, residual errors in the stereo geometry have to be analyzed and corrected electronically during post-production, or using an appropriate pre-calibrated image rectification module suitable for broadcast scenarios. If zoom lenses shall be used, it has to be considered, that a single calibration step is not sufficient as the principal point can vary with the focal length [Wu13]. Consequently, lookup-tables with different correction terms which depend on the focal length have to be pre-calibrated.

5.6.1.2 Proposed Workflow

The proposed workflow involves the use of the components described in section 5.3 and the algorithms for the geometry analysis described in 5.4 in particular.



Figure 5.10. Anaglyph image visualization after the assisted mechanical alignment but before applying the image rectification. Vertical disparities are visible in the anaglyph image.

The stereographer is able to switch between different visualization modes of the stereo images such as anaglyph, or difference image, or half transparent overlay. These overlay images give the stereographer fast visual feedback of the precision of the mechanical calibration process or the image rectification. Alternatively, the stereographer can follow the numerical advises computed by geometry analysis to perform a good mechanical interpretation. Both can be combined to achieve a fast and precise calibration.

Figure 5.10 illustrates the above mentioned features. The red circles highlight the current geometric parameters used for the mechanical alignment. In this example a remaining roll error of 1° is detected beside a small tilt error and a slight mismatch of the two focal lengths, a result from the stereo analysis described in 5.4.2.

Please note the vertical disparities in the upper left corner of the image which are caused by a significant roll of camera and which can easily be recognized in the presented anaglyph overlay mode. If the stereo rig allows adjusting the roll angle, the stereographer can now adjust the respective control unit or screw until the roll error vanishes. The same procedure can be repeated for the other geometrical imperfections.

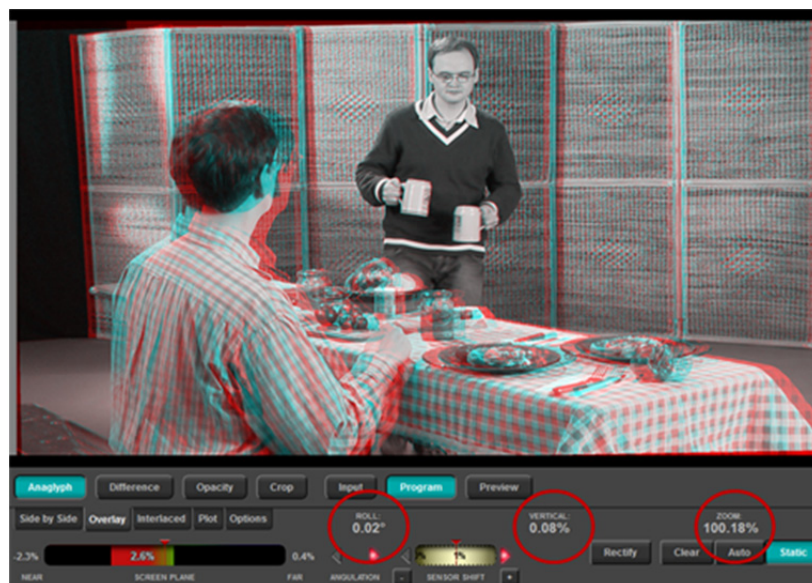


Figure 5.11. Anaglyph image visualization after applying the image rectification.

Remaining geometrical distortions which cannot be compensated by rig adjustments can be eliminated by image rectification. Figure 5.11 illustrates the effect of this kind of electronic correction. Compared to Figure 5.10, geometrical distortions have been minimized considerably. This can be verified by inspection of the upper left corner in the anaglyph overlay image of Figure 5.11.

In the case of a broadcast scenario, the stereo geometry might change dynamically, e.g. due to the use of zoom lenses. Using the temporally consistent stereo image rectification implemented in the system described in sections 5.3.3 and 5.3.4, the corresponding errors can be eliminated dynamically.

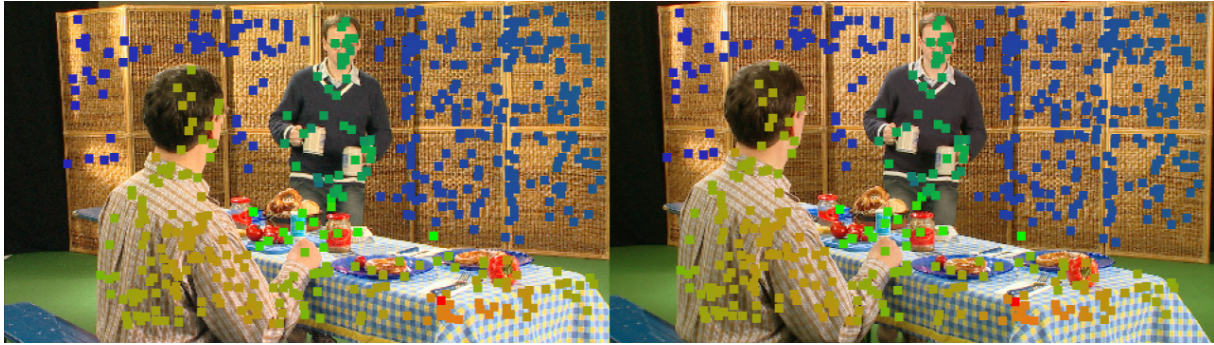


Figure 5.12. *Feature points used for the temporally consistent stereo calibration.*

An important prerequisite for the camera pose estimation are reliable feature points as shown in Figure 5.12. Thus, if no feature points can be detected, e.g. if only a green screen is visible, the proposed approach cannot determine the stereo geometry. Moreover, scene parallax is needed to calibrate for potential translational errors, i.e. feature points in a single plane are not sufficient for a reliable 3D pose estimation.

5.6.1.3 Workflow Enhancements

Both, the legacy and the proposed calibration workflow are suitable to precisely calibrate a stereo rig. The former workflow however, requires much more expertise and know-how and the use of calibration charts which might be unwanted to be used in a set-environment. The proposed workflow in contrast does not involve the use of calibration charts while all geometric errors can be analyzed simultaneously. If all degrees of freedom have motorized axes, the calibration procedure can be performed fully automatic, if required. However, the calibration procedure relies on feature points analyzed from the scene which means that if no feature points can be detected, the calibration process could fail. Consequently as backup a calibration chart could be used to add structure to the scene allowing the generation of sufficient feature points for the geometry analysis process.

In case of broadcast scenarios and the use of zoom lenses, the proposed workflow does not require pre-calibrated lookup-tables which can become complex if effects such as the backlash of zoom motors, the focal length, and the focus position (which might also influence the focal length) have to be taken into account. Moreover, in the case of thermal dilation of components of the stereo rig or unexpected events such as the displacement of the semi-transparent mirror, pre-calibrated lookup tables cannot correct all remaining disparities. Within the proposed approach, residual errors irrespective of its origin can be minimized with a small temporal delay. If very fast changes of the focal length are expected, a combination of both workflows might be useful, although according to [Mediburu12], the focal length shall not be changed too fast for an enjoyable 3D sensation.

5.6.2 Dynamic Convergence Plane Adjustment

5.6.2.1 Manual Convergence Adjustment Workflow

The convergence adjustment is performed by the convergence puller in the case of a manual adjustment. This can be performed by changing the convergence angle of the stereo cameras, or by adjusting the horizontal image translation, e.g. using a remote control connected to an image processing unit. In order to identify the current convergence plane, a constant visual feedback is required.

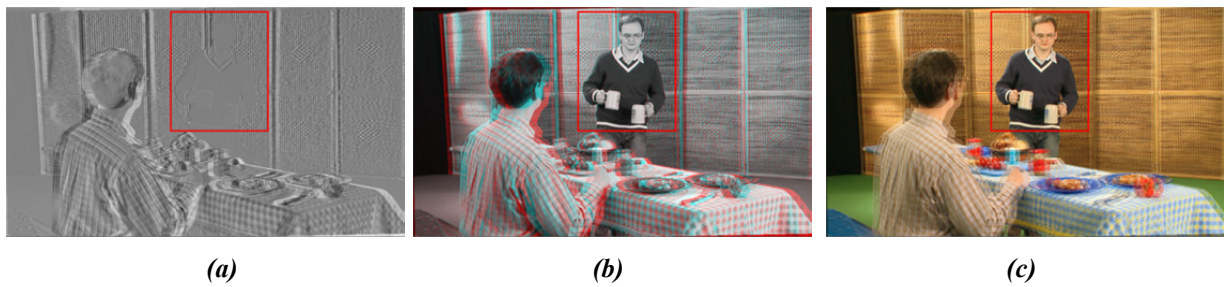


Figure 5.13. *Different visualization modes to monitor the convergence plane or to interactively steer the convergence plane adjustment: (a) Difference of the luminance levels. (b) Anaglyph mode. (c) Overlay of the two input images.*

Figure 5.13 shows different visualization modes which are suitable to identify the current convergence plane which is set to the person in the center of the image inside the red rectangle. In the luminance difference visualization mode from Figure 5.13 (a), the objects in the convergence plane appear as flat gray surface. In the anaglyph mode from Figure 5.13 (b), objects in the convergence plane have sharp edges without red and cyan colored boundaries. In the image overlay visualization from Figure 5.13 (c), the images from the left and right camera are overlaid. Objects in the convergence plane are sharp without a double-image.


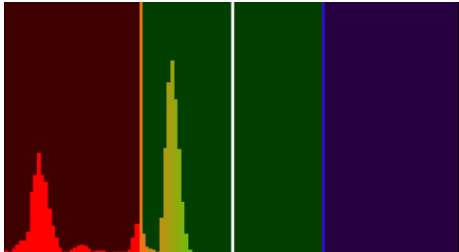
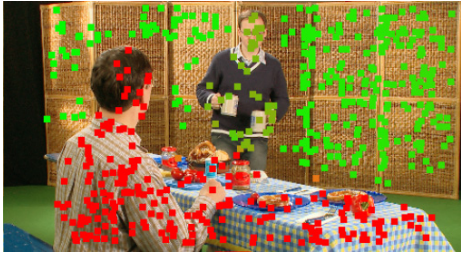
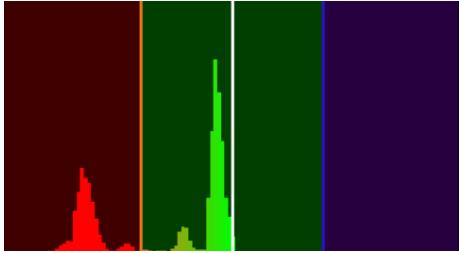

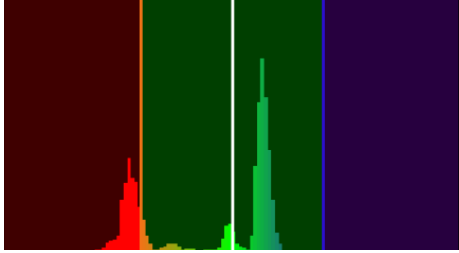
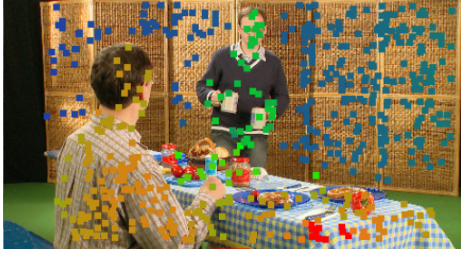
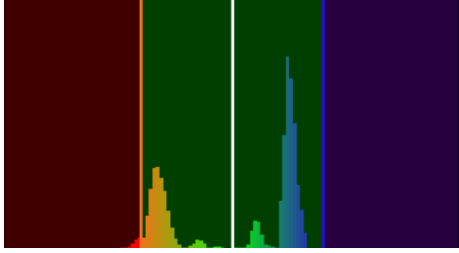
Beside the pure technical requirements, the convergence puller has to address artistic aspects which also influence the choice of a specific convergence plane. Moreover, in dynamic scenes, an anticipation of the movements of objects which shall remain in convergence can be performed by the convergence puller.

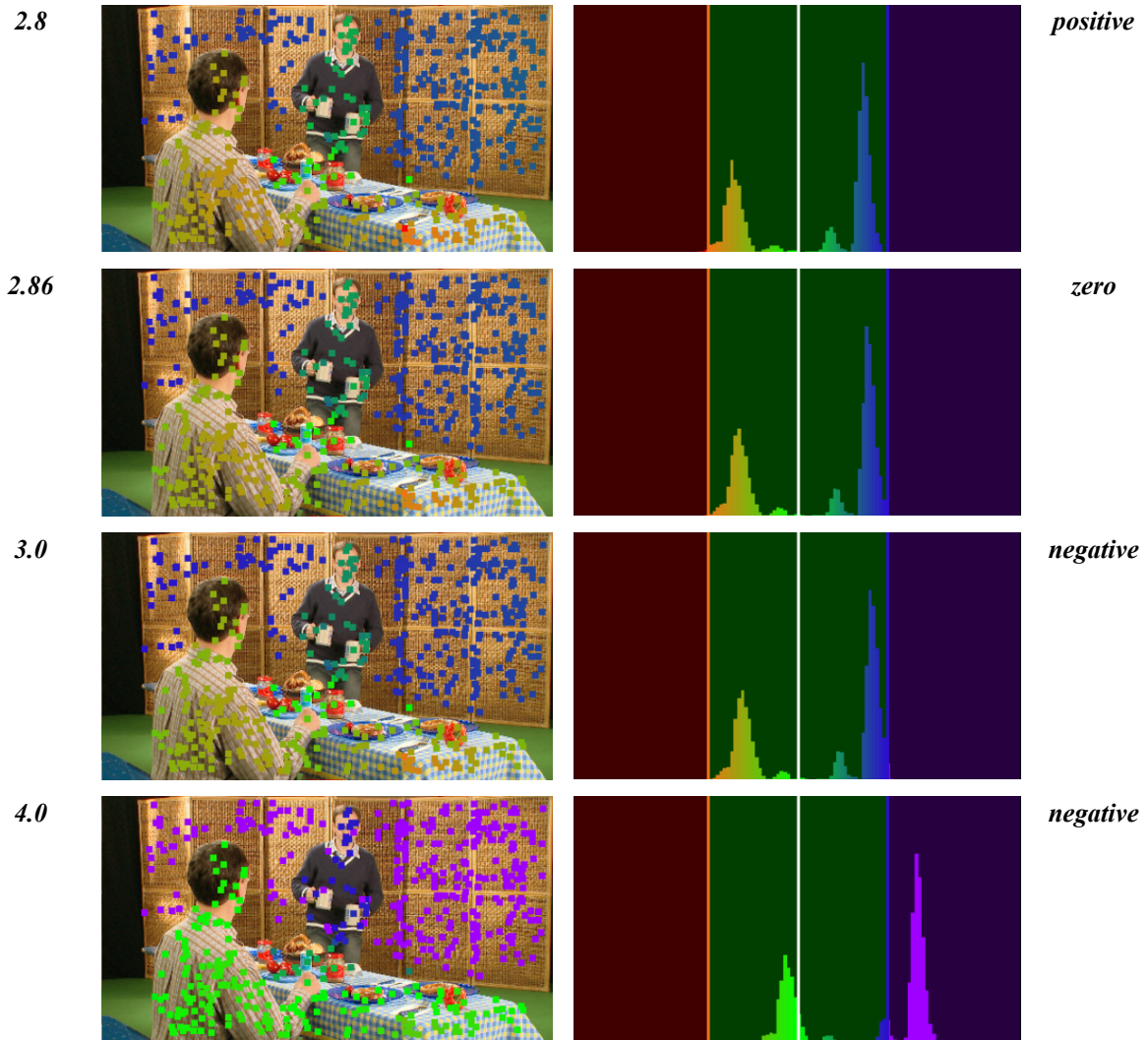
5.6.2.2 Proposed Workflow

It is proposed to apply the algorithm for the automatic derivation of the horizontal image translation from 5.5.2 for the electronic control of the convergence plane. The algorithms can be applied for 3D live applications or in post-production, either in manual or automatic mode. The concept is very similar for all these scenarios. Using the robust feature detection described in section 5.4.1 and the derivation of a disparity histogram, a near and far clipping plane is calculated as described in section 5.5.1. In this example, the limit for the negative and positive parallax are set to $\tilde{\mathcal{P}}_{min} = -2.0$ and $\tilde{\mathcal{P}}_{max} = 2.0$. Typical values for parallax limits depending on the screen size are given in Table 2.1 on page 30. These values are used to derive a new value for the horizontal image translation (HIT), or

more precisely a correction term ΔHIT which describes in which direction the convergence plane has to be adapted as illustrated earlier in Figure 5.5. In the automatic operation mode, the ΔHIT is applied by the assistance system, while in manual operation mode, the convergence puller can follow the advice which is visualized inside the graphical user interface. To monitor the current settings efficiently, the feature points and the disparity histogram are shown in the GUI which are colored according to their disparity. In Table 5.1, this behavior is illustrated.

Table 5.1. Visualization of the automatic derivation of the horizontal image translation HIT . The visualization of the disparity histogram is divided in three segments: the leftmost segment colored in red indicates disparities which are below the minimum parallax value \tilde{P}_{min} , the rightmost segment colored in violet indicates disparities beyond the maximum parallax value \tilde{P}_{max} . The center segment colored in green corresponds to comfortable viewing range. The feature points are colored accordingly with a color gradient ranging from red to violet. The white vertical line in the disparity histogram corresponds to the zero parallax plane.

HIT	Image with Feature Points	Disparity Histogram	ΔHIT
0.0			positive
1.0			positive
2.0			positive
2.6			positive



In the first column, the current HIT value is shown, while in the second and third column, the feature points and the disparity histogram are visualized. Following the algorithms described in section 5.5.2, a correction term for the current HIT denominated as ΔHIT is calculated. The signum of the ΔHIT is shown in the rightmost column. It can be seen that the signum is either positive or negative until the optimal HIT value (here it is reached for HIT=2.86 in row 6) has been found. When starting with a HIT value of 0 as shown in the first row, the HIT is periodically increased through the rows two to five until reaching row six. The similar behaviour is shown when starting in the last row with a HIT value of 4.0. Now, the HIT needs to be decreased, hence the indicator for the ΔHIT is negative until the optimum in row six has been reached.

5.6.2.3 Workflow Enhancements

An advantage of the automatic workflow is surely that personnel can be saved. Instead of requiring one convergence puller per stereo pair, a supervising stereographer could monitor several stereo rigs at the same time. Moreover, the quality of the convergence adjustment remains at the same level even over a long period. Humans in contrast can achieve higher quality results, if well trained, and also take

artistic choices into consideration and anticipate situations in the near future. On the other hand, performing a routine task while keeping a high level of concentration over a long period can become very demanding, especially without breaks and pauses. Finally, the automatic and dynamic convergence plane adjustment relies on the existence of reliable feature points. If the environment inhibits for instance extreme low light conditions, a default convergence plane might be the best choice.

5.6.3 Adjustment of the Inter-axial Distance

5.6.3.1 Legacy Workflow

Different strategies exist to calculate the optimal inter-axial distance or stereo baseline \mathcal{B} . If the distance of the nearest object from the camera Z_{near} and the distance of farthest object are known as well as the focal length f and the depth budget $\tilde{\mathcal{P}}_{max} - \tilde{\mathcal{P}}_{min}$, the inter-axial distance can be calculated according to eqns. (2.10) and (2.12) from section 2.3. The respective distances can thereby be measured using measuring tape or a laser rangefinder.

Alternatively, a manual image based analysis of the depth volume can be performed by measuring the parallax in the stereo image pair.

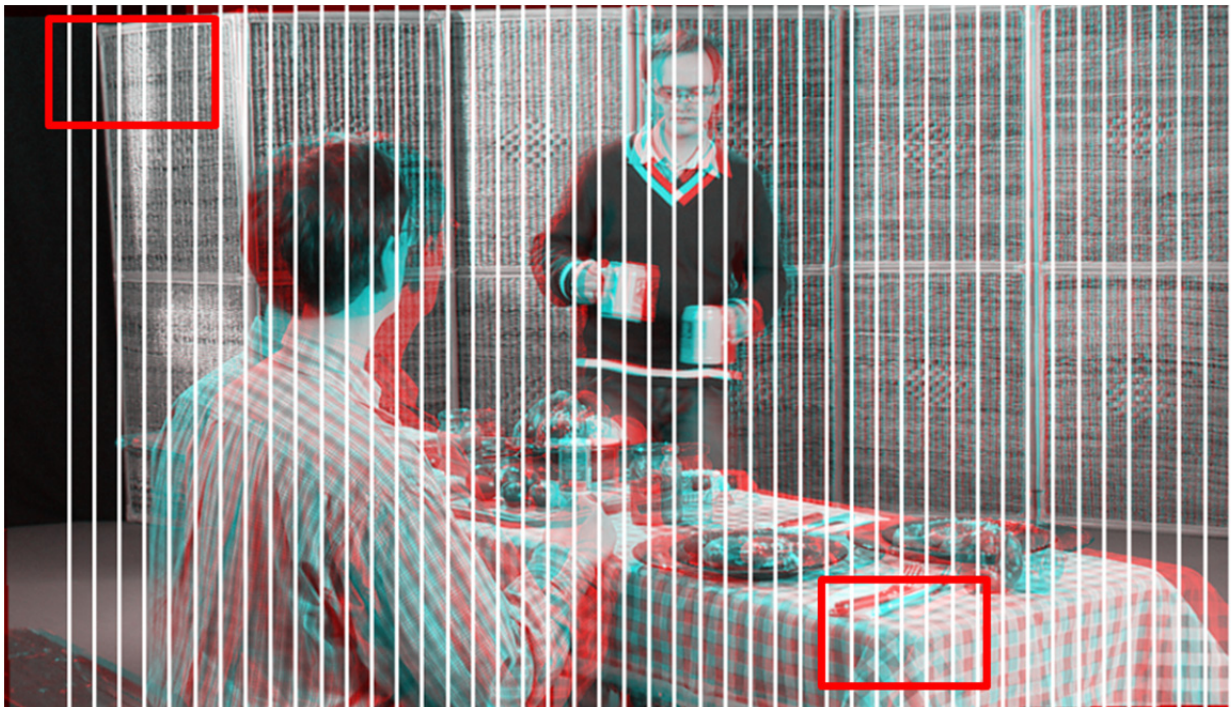


Figure 5.14. Anaglyph image with overlaid vertical grid lines painted in white color. The separation between adjacent grid lines amounts to 2% of the image width. The farthest object and the nearest object are marked by a red rectangle in the upper left corner and the lower right part of the image respectively.

In absence of additional tools, it is possible to measure the parallax inside a stereo image pair using grid lines which are overlaid on an anaglyph image as shown in Figure 5.14. In this figure, the white vertical lines have a separation of 2% of the image width. Now, the farthest visible object in the stereo

pair is searched and the two images are horizontally translated such that the farthest object is in convergence, i.e. has zero disparity. The farthest object in Figure 5.14 is the curtain in the upper left corner of the image. Now, the nearest object in the scene is located to visually examine its horizontal disparity. In the example from Figure 5.14, the nearest object is near the edge of the table in the lower right part of the image. The two regions of interest are marked by a red rectangle in Figure 5.14. A magnified cut-out of these regions is shown in Figure 5.15.

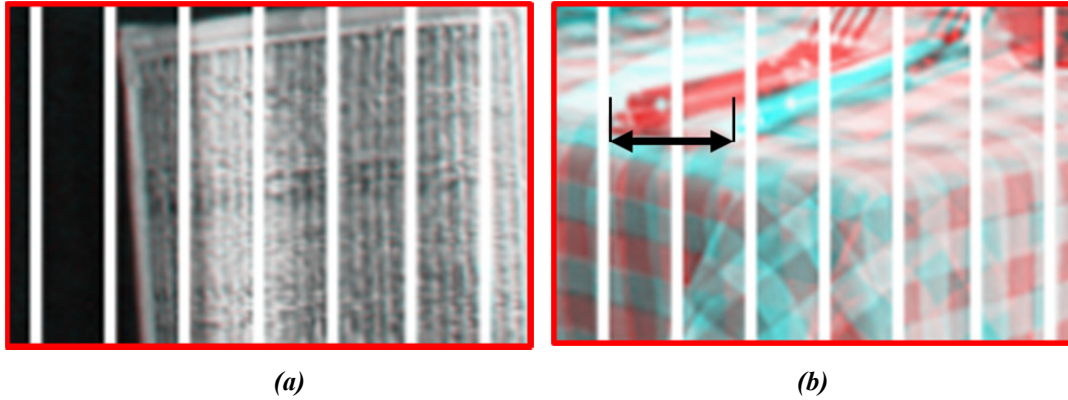


Figure 5.15. Details of the grid lines image with cut-out from Figure 5.14. The horizontal parallax of the background (a) vanishes while the disparity in the foreground (b) is lower than two grid lines, i.e. below 4% of the image width but more than one and a half grid lines, i.e. more than 3% of the image width. The disparity is indicated by the black arrow in (b).

As indicated in the description of Figure 5.15, the manual measurement of the horizontal disparity in Figure 5.15 (b) seems to be between 1.5 and 2.0 grid lines, i.e. around 3-4% of the image width while the horizontal disparity in Figure 5.15 (a) was zeroed out. By applying the nomenclature from section 5.5.1 it can be concluded that the disparity range $disp_{range}$ is the following:

$$3\% < disp_{range} < 4\%. \quad (5.11)$$

Please note that the object used to measure the disparity in Figure 5.15 (b), i.e. the fork, is slightly behind the near clipping plane which lies near the edge of the table in the foreground. However, the pattern of the tablecloth is unsuitable for the disparity measurement due to its repetitive pattern. A higher accuracy of the disparity measurements can be achieved when sharp horizontal edges can be found in the scene. If required, calibration charts as shown in Figure 5.8 can be used. Another source for inaccurate measurements is shown in Figure 5.16.

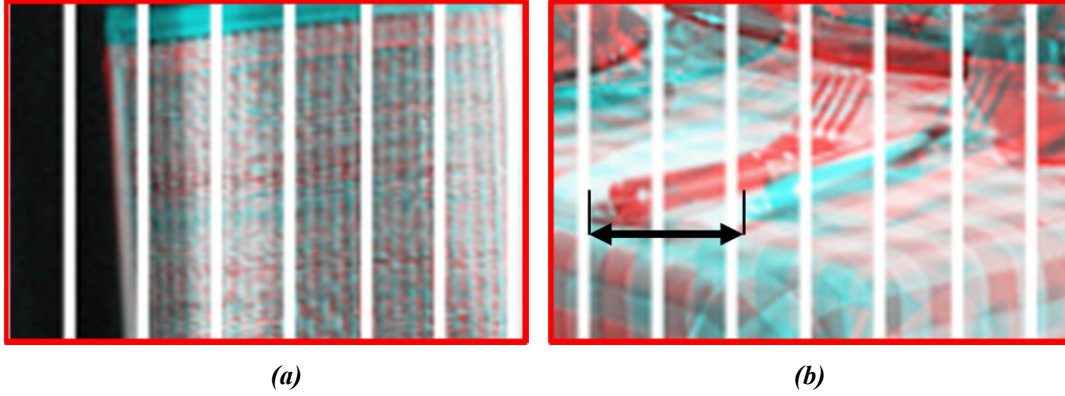


Figure 5.16. *Measuring the horizontal disparity in the presence of vertical disparities can lead to inaccurate results. Compared to Figure 5.15 (b), the horizontal disparity in (b) is much higher.*

Grid lines have been applied to an unrectified stereo pair in Figure 5.16. This results in a higher horizontal disparity measured in Figure 5.16 (b) which is higher than 2.0 grid lines and lower than 2.5 grid lines, i.e., given that the separation between two grid lines stands for 2% of the image width, the measured disparity is:

$$4\% < \text{disp}_{\text{range}} < 5\% \quad (5.12)$$

By comparing equations (5.11) and (5.12), one can observe that the disparity measured using the unrectified stereo pair is significantly higher than the one measured using the rectified or aligned stereo pair.

5.6.3.2 Proposed Workflow

It is proposed to use Algorithm 5.2 for the depth bracket analysis from section 5.5.3 to derive the optimal inter-axial distance in combination with a graphical user-interface for the interaction with the stereographer in order to allow for a guided adjustment of the inter-axial distance. Please note that the algorithm uses geometrically corrected feature points for the calculation of the disparity histogram from Table 5.1 and can therefore cope with unrectified input images. The feature points shown in Table 5.1 also indicate that feature points have been identified at both, the near clipping plane and the far clipping plane. Important settings such as the minimum and maximum allowed parallax with respect to different screen sizes can be controlled via the graphical user interface. In the sample from Table 5.1, the limits for the minimum and maximum parallax were set to -2% and +2% of the screen width respectively. Typical settings which take the screen size into account were given in Table 2.1 on page 30 which originates from [Knorr12].

By applying the proposed workflow to the samples from Figure 5.15 and Figure 5.16, the values shown in Table 5.2 were extracted and visualized.

Table 5.2. *Parallax measured from the same scene using the rectified and the unrectified stereo pairs. Using the unrectified stereo pair, the parallax is overestimated.*

Image Pair	Measured Parallax
Rectified Stereo Pair	3.7 %
Unrectified Stereo Pair	4.4 %

The values from Table 5.2 are consistent with the manual measurements from eqns. (5.11) and (5.12). Please note that the system is able to cope with unrectified data, so if required the measured parallax value corresponding to the rectified stereo pair can be calculated. Given the data from Table 5.2, the inter-axial distance can now be updated by calculating a corresponding update ΔB according to eqn. (5.10) and Figure 5.6. The repositioning of the inter-axial distance can be performed automatically, if the corresponding axis is motorized, or manually.

Beside the main purpose of the algorithm to steer the inter-axial distance during the calibration phase or the shooting itself, another use-case for the algorithm is the post-production. Although in post-production the inter-axial cannot be changed anymore, the result of the analysis can still help to identify shots which fit into a given depth budget and to get an objective quality metric for stereoscopic 3D content.

5.6.3.3 Workflow Enhancements

The proposed workflow allows a precise and dynamic measurement of the depth volume without calibration charts or meta-data such as the focal length. It is also able to cope with unrectified image data. As a result, a significant reduction of the setup time and increase of the 3D quality can be expected.

However, it relies on feature points, hence if the scene is not suitable for feature detection and matching, e.g. if only a single green screen without detectable feature points is visible, the geometry cannot be analyzed and hence only a default inter-axial distance could be chosen. Consequently, as a backup, alternative measurement possibilities such as the grid lines from section 5.6.3.1 should be integrated in the graphical user-interface from section 5.3.7.

5.7 Conclusion

A technical description of the assistance system was given in section 5.3. The underlying algorithms were explained in sections 5.4 and 5.5. A comparison of new workflows which were enabled using the new algorithms was performed with respect to existing workflows in section 5.6. In addition, different test productions and field trials have been carried out to evaluate the assistance system under real working conditions. An overview of the test productions is given in the appendix in section 8.1. The feedback gathered during these productions was used to continuously improve the system. The possible workflow enhancements were discussed in the respective sub-sections 5.6.1.3, 5.6.2.3, and

5.6.3.3. As summary, the main improvements consist of a faster and more precise calibration of the mechanical setup and control of the stereoscopic parameters such as convergence plane and inter-axial distance. The analysis is purely image-based and does not require special meta-data or pre-calibrated look-up tables. Analysis results can also be used for post-production purposes.

Although the image-based analysis of the stereo geometry has different favorable aspects, it has also some limitations. It relies on the accuracy and robustness of the underlying feature detectors. The feature detector presented in chapter 4 has shown to be robust against noise and outliers. However, for nearly every image processing algorithm, it is possible to construct input image scenarios, where the algorithm fails. In the case of the stereoscopic 3D production, such a scenario is for instance given if the full field of view is covered by an unstructured homogeneous area such as a green screen, or if the studio or shooting environment is too dark for reliable feature detection. Although the problems addressed by the proposed algorithms are important for a proper 3D production, it has to be mentioned that additional challenges remain, namely a proper setting of the focus and an equalization of the luminance and chrominance settings of the two cameras which could be added to the assistance system in the context of future work.

6 Mixed Baseline Stereo Estimation

6.1 Introductory Remarks

Within this chapter, a new algorithm for the mixed-baseline disparity estimation suitable for depth-image-based-rendering applications is proposed¹⁹. The development of the algorithms was performed in the context of a new multi-camera content-acquisition workflow which was investigated within the European research project MUSCADE [Muscade]. This workflow involves four cameras including a standard stereo camera system as narrow baseline and targets different types of displays such as auto-stereoscopic displays and light-field displays.

Although the main novelty presented within this chapter is the stratified multi-baseline disparity estimator, the workflow takes advantage of many components proposed within this thesis. The calibration of the multi-camera rig and the multi-camera rectification is based on components presented in chapters 3, 4 and 5. This comprises a multi-camera assistance system based on the stereo assistance system from chapter 5 which analyzes the multi-camera geometry using the feature descriptor SKB from chapter 4, and the trifocal tensor estimation method proposed in chapter 3.

Please note that an extensive overview of related work regarding multi-camera content generation was given in section 1.2.3 while theoretical concepts of the disparity estimation process were introduced in section 2.6.

The remainder of this chapter is structured as follows: In section 6.2, the mixed baseline multi-camera geometry and the components used within the MUSCADE setup, i.e. the beam-splitter and its extension towards four cameras along with necessary pre-processing steps such as calibration are described. Section 6.3 is the main section where the novelty in this chapter, i.e. the stratified generation of the multi-view video plus depth (MVD) content is described in detail, from initial pairwise disparity estimation, to merging and post-processing of the disparity maps. In section 6.4, the quality of the MVD content and the suitability of the inner stereo pair for glasses based stereoscopic 3D applications will be evaluated before concluding the chapter in section 6.5.

6.2 Multi-Camera Content Acquisition and Pre-Processing

6.2.1 The MUSCADE Multi-Camera Setup

The multi-camera rig used in the MUSCADE test production consists of four identical cameras in linear camera array configuration. Figure 6.1 gives an overview of the positioning of the four cameras on the multi-camera rig. As typical for linear camera arrays, all four optical centers are aligned on a common baseline. This simplifies the production workflow as all cameras can be rectified jointly and

¹⁹ Parts of the content in this chapter have been previously published in [Zilly12b], [Zilly2013] and [Zilly14].

all subsequent processing steps such as depth estimation and Depth Image Based Rendering (DIBR) can be conducted in a line-wise manner which simplifies the parallelization of the underlying algorithms and thus eases to meet the real-time processing constraint.

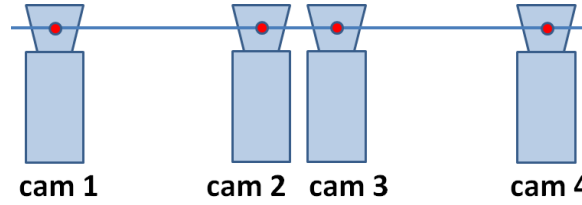


Figure 6.1. Schematic drawing of the four-camera rig. All cameras' optical centers are aligned on a common baseline although the arrangement of the cameras is non-equidistant.

To ensure the proper alignment and positioning of the cameras, a dedicated camera assistance system as described in chapter 5 and [Zilly10b] was used. The two cameras in the center (i.e. cameras 2 and 3 from Figure 6.1) form a narrow stereo baseline system. Beside depth estimation purposes, these two cameras are used to capture content suitable for native displaying on stereoscopic displays, i.e. without the need to involve DIBR. But the arrangement in a narrow baseline has also additional benefits. The depth estimation process is more robust and produces denser depth maps compared to a wide baseline setup as the parallax and associated occlusions are quite small.

The two outer cameras, also referred as satellite cameras (i.e. cameras 1 and 4 from Figure 6.1), constitute the wide baseline system. The high amount of parallax inherent to this system allows the DIBR process to interpolate a high number of views as required for today's light-field displays without the need to extrapolate from the stereo baseline. View extrapolation artifacts arising from growing disocclusions can be reduced using methods such as smoothing the depth maps prior to the DIBR [Fehn03]. However, these methods work only within certain limits because missing information can only be *guessed*. In contrast, when synthesizing views between the cameras 1 and 4, there are always one or more views left and right to the virtual view. Hence, parts of the image which might be occluded in camera 1 — and disoccluded when generating the virtual view — can most probably be filled with information visible in camera 4. Apart from the additional amount of parallax, the wide baseline system enhances also the precision of the depth maps, i.e. a higher depth resolution can be achieved. In fact, disparities estimated between camera 1 and camera 2 will be proportional to the respective disparities estimated between camera 2 and camera 3 by an amount equal to the ratio of the camera baselines.

It was aimed in the MUSCADE project to use high quality HD-TV-cameras and lenses. However, the camera bodies and diameters of professional grade lenses can easily become as large as 15 cm. Consequently, a standard side-by-side configuration of the cameras was not an option as conforming to existing 3D production rules [Mendiburu08, Zilly11a] a typical inter-axial distance, i.e. the

separation of the cameras' optical axes, of 3-7 cm is required to keep the parallax reproduced on the 3D display within a comfortable viewing range (cf. section 2.3.1).



Figure 6.2. *Multi-camera rig with narrow and wide baseline. The inner two cameras are mounted on a beam-splitter.*

Hence, to bring the optical axes near enough, the two narrow baseline cameras are mounted on a beam-splitter as described in the introductory section 1.2.1. The two wide baseline cameras are mounted outside the mirror box. A picture of the fully equipped multi-camera rig is shown in Figure 6.2.

6.2.2 Calibration of the Linear Camera Array

The satellite cameras outside the mirror box are attached onto adjustable mounting plates which allow performing a precise mechanical alignment. The architecture of the multi-camera rig along with calibration plates for the precise positioning of the cameras was jointly developed within the MUSCADE consortium. A multi-camera rectification onto a common baseline facilitates further processing steps such as Depth image Based Rendering (DIBR). In fact, DIBR reduces to a simple horizontal pixel shifting as described in section 1.2.2.2, as the virtual camera to be rendered will also lie on the common baseline and only horizontal parallax will occur. This means that the rendering process can easily be parallelized and simple 1D heuristics can be applied for interpolation and hole filling operations. The alignment process was conducted using the camera assistance system described in chapter 5.

6.2.3 Multi-Camera Rectification

The multi-camera rectification algorithm which was developed for the MUSCADE approach is based on the estimation of the trifocal tensor from chapter 3 and extended towards four cameras by defining two camera triplets and estimating the corresponding trifocal tensor twice. The left triplet is composed of the inner stereo pair and the left satellite camera (cameras 1, 2, and 3), while the right triplet consists of the inner stereo pair and the right satellite camera (cameras 2, 3, and 4). The left camera from the inner stereo pair (i.e. camera 2) is the so-called hero view which remains unchanged. It serves

as common reference camera and remains unaffected by the rectification process. In a first step, the trifocal tensor for the right camera triplet is estimated yielding to rectifying homographies for cameras 3 and 4, while camera 2 is kept untouched. Subsequently, the trifocal tensor for the left camera triplet is estimated yielding to a rectifying homography for camera 1 while taking into account the already estimated geometry of camera 3, i.e. cameras 2 and 3 remain untouched.

The full mathematical derivation of the rectification algorithm along with an evaluation of the residual back-projection error of the multi-camera rectification algorithm and a strategy to reduce the linearization error through iteration was given in chapter 3. For the remainder of the chapter it is assumed that required rectifying homographies and the ratios β_{12} , β_{23} , and β_{34} of the camera baselines are known where β_{12} denotes the baseline between cameras 1 and 2, while β_{23} , and β_{34} denote the baselines between the cameras 2 and 3 and the cameras 3 and 4 respectively.

6.3 Stratified Mixed-Baseline Disparity Estimation

In this section, details of the proposed mixed-baseline disparity estimation algorithm are given. The algorithm is thereby inspired by the multi-view disparity estimation approach proposed in [Hirschmüller08]. Instead of creating a joint similarity measure between corresponding pixels using all available views as proposed in [Okutomi93], the approach of [Hirschmüller08] is followed which suggests to perform a stereo disparity estimation of neighboring views which shall be merged subsequently. The latter approach is more suitable when large disocclusions in the stereo views are apparent.

6.3.1 Mixed Baseline Stereo Setup

Figure 6.3 illustrates the implications of the mixed narrow and wide baseline. Camera 1 corresponds to the left satellite camera, while cameras 2 and 3 correspond to the inner stereo pair with narrow baseline. Camera 4 has been omitted in this illustration as no additional insights would be generated. In fact, the effect of the right wide baseline is similar to the left wide baseline.

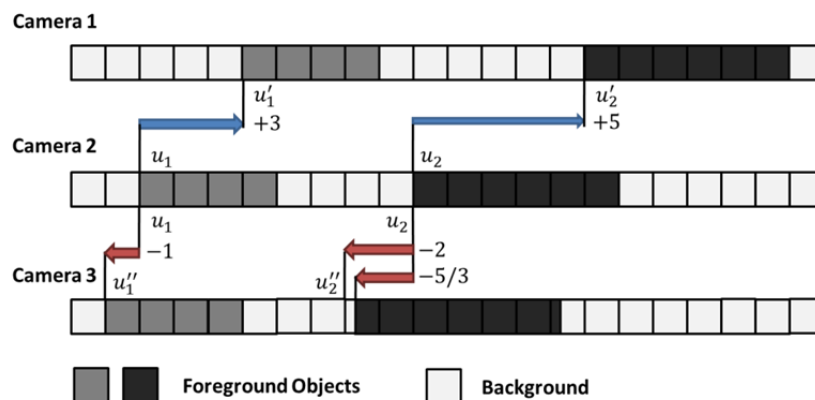


Figure 6.3. Schema of mixed narrow and wide baseline setup. The horizontal disparities are proportional to each other. In this example, the baseline ratio is $\beta = 3.0$. Using the wider baseline involving cameras 1 and 2, a more precise disparity measurement is possible.

Figure 6.3 illustrates schematically the displacement of two objects among the three cameras. In the example, the disparities are proportional to each other with a baseline ratio of $\beta = 3.0$ according to eq. (3.24), i.e. the pixel position of an object in the third camera can be derived, once its position in two other cameras is known. Moreover, the example shows that this displacement can also result in sub-pixel disparities. In the narrow baseline (here between cameras 2 and 3), a naïve disparity estimator might measure a displacement of 2 pixels. Using the wide baseline, a displacement of 5 pixels would be measured. Hence, the accuracy of the disparity measurement is increased as the normalized disparity value retrieved from the wide baseline $\text{disp}_{\text{wide}} = 5/\beta = 5/3$ is more precise than the disparity value estimated using the narrow baseline $\text{disp}_{\text{narrow}} = 2$.

The examples illustrates that wide baseline stereo pairs will produce disparity maps with higher depth resolution. However, the disparity maps are usually sparser when using the wide baseline, while the narrow baseline stereo pair will produce denser depth maps as the similarity between pixel blocks is higher and the amount of occlusions lower. In Figure 6.4 the workflow diagram of the proposed MVD4 generation algorithm is presented which takes into account the mixed baseline geometry. The nomenclature for the disparity maps in Figure 6.4 and this chapter is as follows: **Disp xy** denotes a disparity map estimated for a camera at position x forming a stereo pair with camera y . Moreover, **Disp $x \rightarrow y$** denotes a disparity map which originates from camera position x which has been rendered to the camera position y using DIBR. Finally, **Disp x** denotes the final disparity map for camera x .

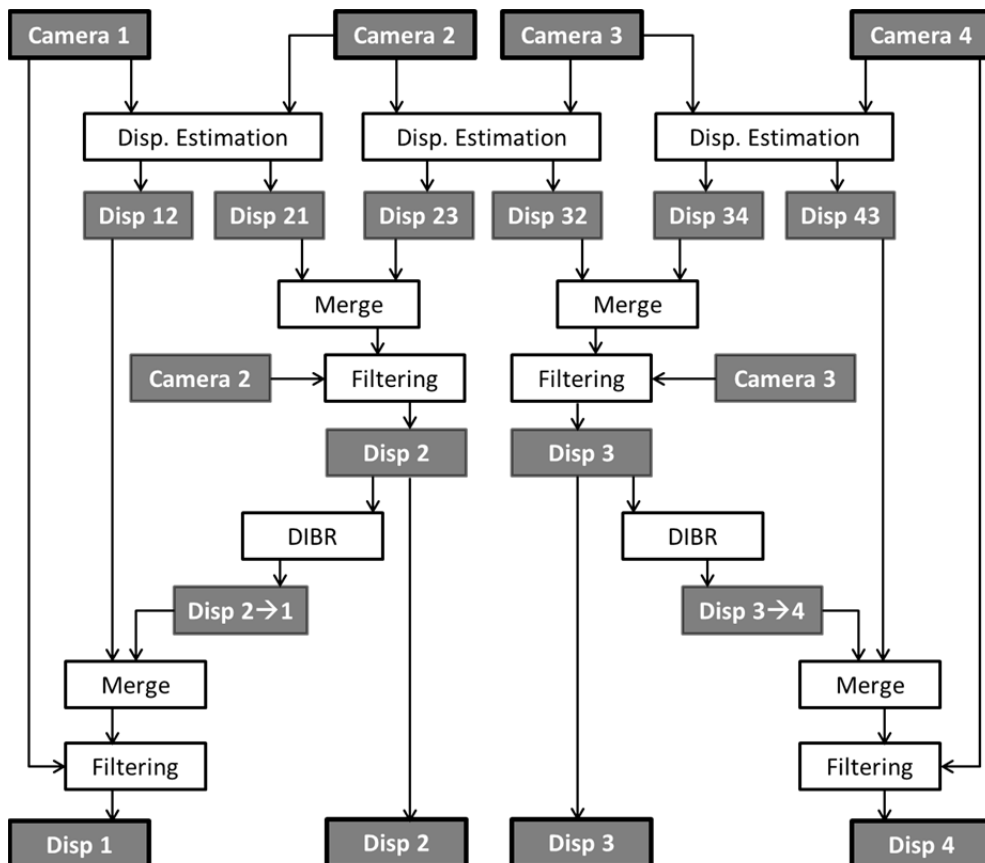


Figure 6.4. Schematic drawing of the proposed stratified mixed-baseline disparity estimation algorithm.

The MVD4 generation process can be stratified in up to six main steps which will be described in the sub-sections 4.1 to 4.6. The stratification allows for a parallelization of the algorithms in different PCs. The four camera images (see Figure 6.16) form three stereo pairs while the two inner cameras are involved in two pairs each. After initial disparity estimation, the disparity maps for the inner cameras are merged and filtered using a cross-bilateral filter resulting in dense disparity maps. Subsequently, these disparity maps are rendered to the outer camera positions. The former transfer can be achieved by applying a DIBR algorithm to the disparity maps. However, although already quite dense, the raw depth maps from the center pair contain still many holes and empty areas which significantly impair the quality when applying DIBR to these depth maps. Better results can be achieved when applying a cross-bilateral filter to the depth maps before the DIBR step. They can subsequently be merged with the initial disparity maps from the satellite cameras before a cross-bilateral filter is applied to the two satellite cameras as well. It results a combination of elementary processing steps such as initial disparity estimation, merging, filtering and rendering of the disparity maps which need to be arranged in the right order. In the following each processing step will be described in more detail.

6.3.2 Initial Disparity Estimation

Given the four camera images (see Figure 6.16) which have been rectified onto a common baseline, three pairs of disparity maps are estimated using the real-time capable disparity estimator HRM [Atzpadin04] along with a left-right consistency check. Please note that alternative disparity estimators such as [Forstmann04] would also be suitable for this task. Different disparity estimators and related work was presented in section 2.6.1. However the HRM is able to estimate temporally consistent disparity maps, which is an important feature for the MVD generation process. For the two inner cameras, two disparity maps are computed, one estimated with the narrow baseline neighbor and one with the wide baseline neighbor, e.g. for camera 2, the disparity maps **Disp21** and **Disp23** (see Figure 6.4 and Figure 6.6) result from the estimation process.

6.3.3 Left-Right Consistency Check

As a result of the initial disparity estimation, two disparity maps for each camera of the inner stereo pair are computed. The first disparity map is estimated with respect to the narrow baseline neighbor while the second disparity map is estimated using the wide baseline neighbor, e.g. for camera 2, the two disparity maps **Disp21** and **Disp23** (see Figure 6.4 and Figure 6.6) are computed. The result of the initial disparity estimation are complementary pairs of disparity maps, e.g. one from camera 1 with disparities pointing to the pixel positions in camera two (**Initial Disp 12** according to Figure 6.5) and one from camera 2 pointing towards camera 1 (**Initial Disp 21**). As shown in Figure 6.5 the corresponding disparities have opposite sign in the two disparity maps.

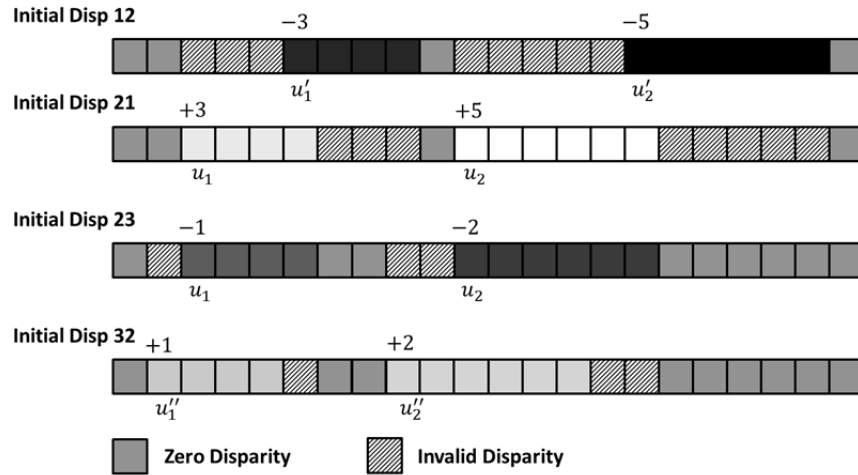


Figure 6.5. After the initial disparity estimation step, positive and negative disparities occur. The absolute value of the disparities is proportional to the respective camera baselines. A left-right consistency check is performed to filter out inconsistent disparities. For instance, the corresponding and consistent pixel from u'_2 (value = -5) in Initial Disp 12 is the pixel at the position u_2 (value = $+5$) in Initial Disp 21.

The left-right consistency check is a suitable method to eliminate wrong disparity values which correspond to pixels which are occluded in one of the two cameras (cf. section 2.6.4). Fortunately, for the inner stereo pair, each camera has a left and a right neighbor. Therefore, there is a portion of pixels which might be occluded in the left neighbor, but visible in the right neighbor. When merging the disparity maps during a processing step described below, at least some of the holes which occurred due to occlusions can be filled.

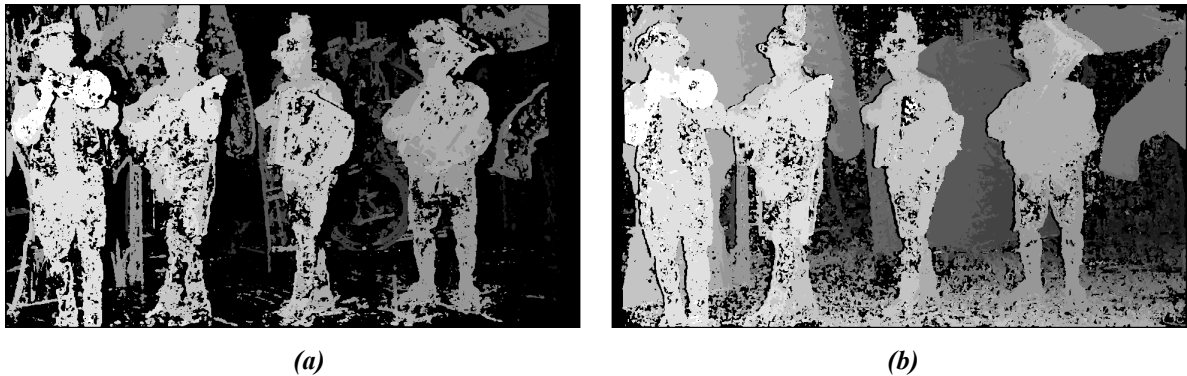


Figure 6.6. Disparity maps estimated for camera 2. (a) From wide baseline Disp21. (b) From narrow baseline Disp23. Please note that in order to expose possible artifacts, a nonlinear rescaling has been applied to the illustrations of the disparity maps in this section 6.3. To allow for a better comparison of the disparity maps, the normalization step described in section 6.3.4 has been applied prior to the nonlinear rescaling.

Figure 6.6 shows the two disparity maps belonging to camera 2. As expected, the disparity map **Disp21** which has been estimated using camera 2 and camera 1, i.e. the wide baseline, is sparser than the disparity map **Disp23** estimated using camera 2 and camera 3, i.e. the narrow baseline.

6.3.4 Normalization of the Disparities

As the cameras' optical axes are parallel after rectification, the disparities are proportional according to the camera baselines (cf. section 0). Hence, by dividing all disparity values by the respective

baseline factors β provided by the multi-camera rectification algorithm, all six disparity maps can be normalized into an inverse-of-a-distance representation as used in [Okutomi93].

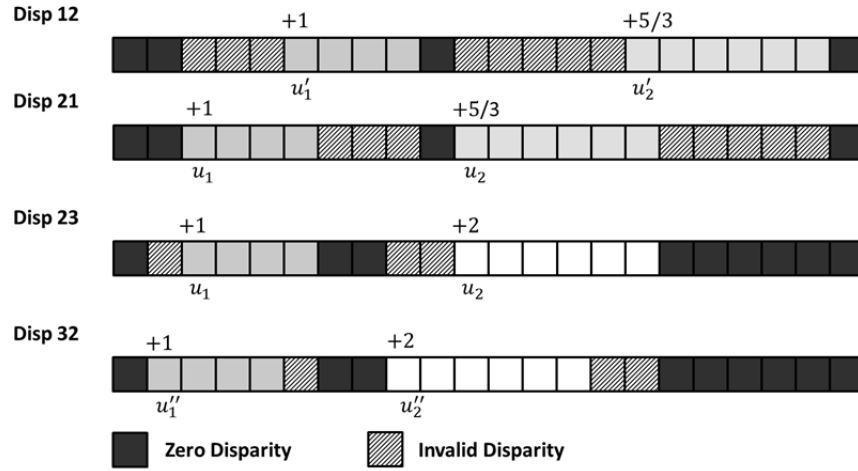


Figure 6.7. Normalization of the disparity maps. All disparities are divided by a value proportional to the respective stereo baseline resulting in an inverse-of-a-distance representation with sub-pixel accuracy

By applying the normalization procedure defined in eq. (6.1), the disparities shown in Figure 6.5 are transformed into normalized disparities as shown in Figure 6.7. Please note that as shown in Figure 6.3 and Figure 6.7, non-integer values (e.g. “5/3”) can occur as a result of the normalization process.

$$\forall i \in [1, \text{size}(\text{Disp}_{12})] : \text{Disp}_{12}(i) = \begin{cases} |\text{InitialDisp}_{12}(i)|/\beta_{12} & \text{if } \text{InitialDisp}_{12}(i) \text{ valid} \\ -1 & \text{(or invalid) else} \end{cases} \quad (6.1)$$

Please note that a 10 bit fixed-point representation of the disparity values is used to allow quarter-pixel accuracy in a range from 0 to 255 within the normalization and all subsequent processing steps. Higher disparity values are not likely to occur as the disparities have been normalized with respect to the inner stereo pair which is formed by a narrow baseline.

6.3.5 Merging of the Inner Disparity Maps

After normalization, the disparity maps can be merged, e.g. **Disp21** with **Disp23** and **Disp32** with **Disp34** (see also Figure 6.4). The process is illustrated in Figure 6.8. The corresponding mathematical operation is defined in eqn. (6.2):

$$\forall i \in [1, \text{size}(\text{Disp}_{21})] : d_{21} = \text{Disp}_{21}(i), d_{23} = \text{Disp}_{23}(i),$$

$$\text{Disp}_2(i) = \begin{cases} d_{21} & \text{if } (d_{21} \geq 0) \wedge (d_{23} < 0) \\ d_{23} & \text{if } (d_{21} < 0) \wedge (d_{23} \geq 0) \\ (d_{21} + d_{23})/2 & \text{if } (d_{21} \geq 0) \wedge (d_{23} \geq 0) \wedge (|d_{21} - d_{23}| < \theta_{tri}) \\ \text{invalid} & \text{else} \end{cases} \quad (6.2)$$

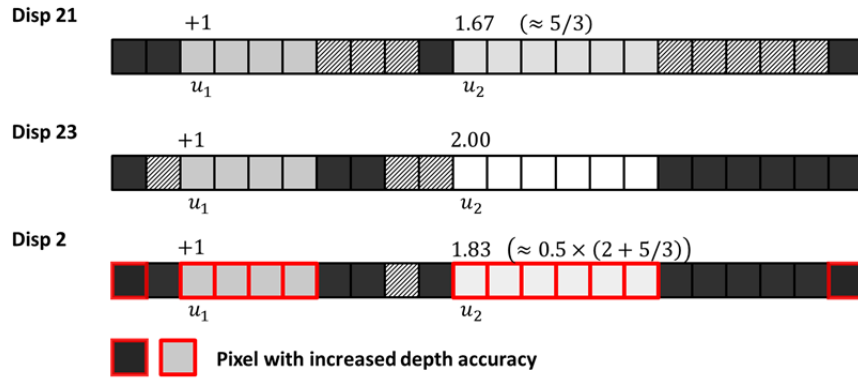
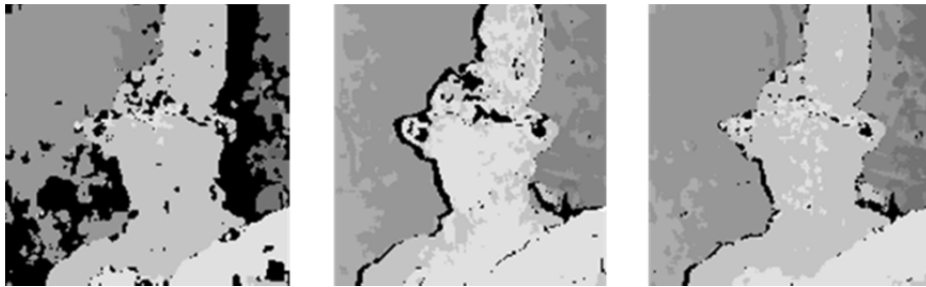


Figure 6.8. Merge process of the disparity maps. The resulting disparity is the average of the two input disparities provided that the difference is below the trifocal consistency threshold θ_{tri} . Otherwise the corresponding disparity is marked as invalid. Trifocal consistent disparities are marked in red.

Disparities which are inconsistent along the two disparity maps are removed. This is equivalent to a trifocal or multi-focal consistency check. Different techniques are known in the literature to take advantage of the multi-camera geometry. However, the approaches proposed in [Okutomi93] and [Campbell08] make only sense with a large number of views. Moreover, within [Okutomi93], it is proposed to sum up the total score of the cross correlation tests. This can prevent the algorithm to be executed in a stratified or parallelized way. Moreover, when correlations are summed up in regions which are partially occluded in some cameras, this can lead to unsatisfactory results [Hirschmüller08]. Instead, as proposed by Hirschmüller the disparities are fused or averaged which are consistent in both disparity maps. This avoids or at least minimizes artificial depth discontinuities at the borders of trifocal consistent image regions. Figure 6.9 illustrates the effect of the merging process using the disparity maps **Disp21** and **Disp23**. The result (Figure 6.9, right) is denser than the two input disparity maps. Ideally, the occlusions from the left and the right object borders are filled.



*Figure 6.9. Cut-out of the normalized disparity maps **Disp21** (left) and **Disp23** (center) used to perform the merge (right).*

Besides increasing the density of the disparity maps and eliminating trifocal inconsistencies, the depth resolution of the inner stereo pair is increased. Figure 6.10 shows the regions within the inner stereo pair whose disparity values have survived the trifocal consistency check. After averaging the depth data with the values from the wide baseline system, these regions have increased depth accuracy.

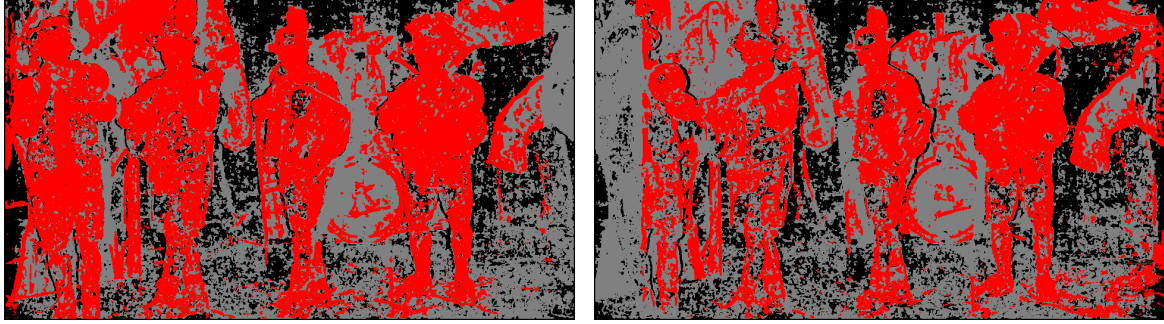
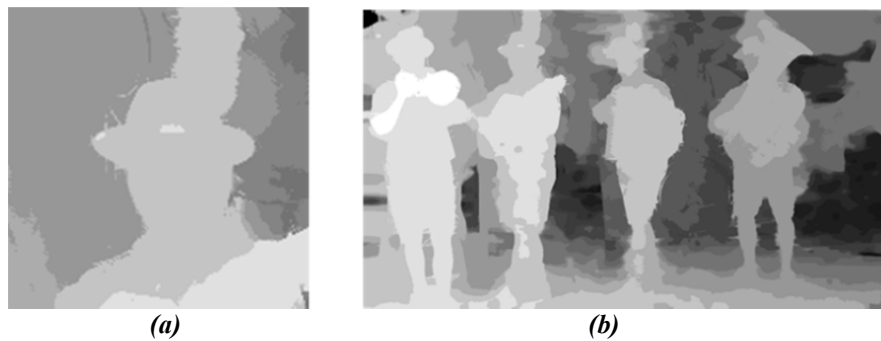


Figure 6.10. Trifocal consistent depth values with increased depth accuracy after the merging process are shown in red.

The foreground objects in the scene can take most advantage of the trifocal consistency. Please note that the higher depth accuracy helps also to give objects rendered by means of DIBR a more detailed depth structure. In contrast, objects with for instance only one disparity value will appear flat or like a cardboard when watched in 3D (cardboard effect) [Mendiburu08, Mendiburu12, Zilly11b]. After the merge, a post-processing filter is applied to the disparity maps as described in the next sub-sections, which will eliminate remaining small depth discontinuities arising from the merge process.

6.3.6 Filtering Inner Disparity Maps

After the merge, a post-processing filter is applied to the disparity which shall eliminate remaining small depth discontinuities arising from the merge process. A cross-bilateral median filter [MüllerM10] which is an extension of the concept described in [Riemens09] and [Kopf07] is applied to the merged disparity maps. The filter is able to fill remaining holes in the disparity map which occurred due to parallax-induced occlusions or inconsistencies of the disparity maps. One could wonder why it is useful to merge the disparity maps given that a cross-bilateral filter can also fill the holes in the disparity maps. In fact, the merged disparity maps have a higher quality, i.e. they are denser in general and contain less outliers. In addition, the depth resolution is increased. All these properties add quality to the filtered result as the bilateral filter has less pixels to fill (which means guessing the disparity) and more reliable and accurate pixels as data basis. Figure 6.11 shows the resulting filtered disparity maps *Disp2* (see also Figure 6.4).



*Figure 6.11. Disparity map *Disp2*. (a) Cut-out from *Disp2*. (b) Disparity map *Disp2* filtered using cross-bilateral median filter.*

As can be seen, after filtering, no holes are left in the disparity map, i.e. it is pixel dense, and the alignment with the object borders has been greatly improved. Moreover, the disparity maps appear

now smoother than before. In fact, many high frequencies which are not near depth discontinuities have been removed. This property along with the similarity of the data after the normalization step enhances also the coding efficiency of 3D video coding [ZhangQ11].

6.3.7 DIBR of Disparity Maps from Inner to Outer Views

After filtering the disparity maps of the inner views, they can be rendered to the viewing position of the satellite cameras by applying the depth-image-based rendering mechanism [Fehn03a] to the disparity maps itself. The reasons for doing this are twofold. On one hand, the initial disparity maps for the outer views are quite sparse compared to the inner views as the parallax is higher and less reliable matches can be found. On the other hand, a trifocal consistency check can be applied as each disparity map from the inner views incorporates information from both central views.

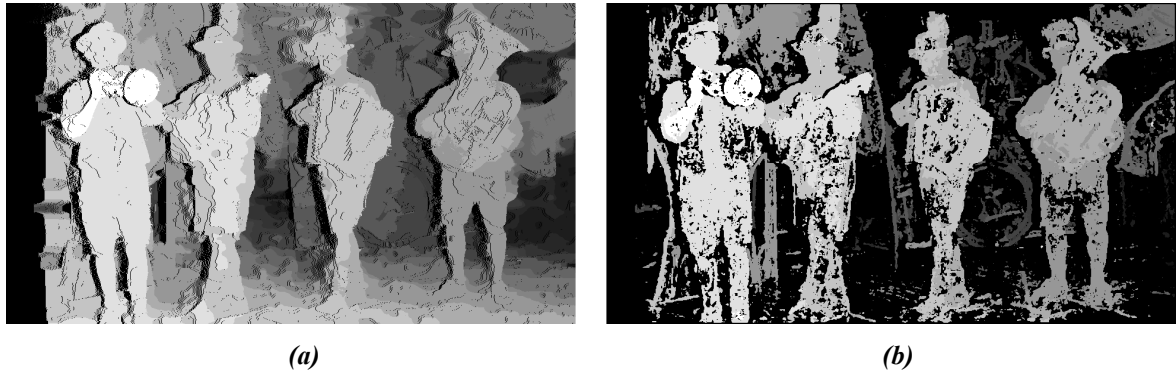


Figure 6.12. (a) Result of the DIBR transfer of $Disp_2$ to camera 1 $Disp_2 \rightarrow 1$. (b) Sparse disparity map $Disp_{12}$.

The rendering of the disparity maps follows a simple algorithm. Each pixel is shifted by an amount proportional to its disparity value of the respective wide camera baseline. A simple forward mapping with z-test is performed according to eqn. (6.3) and Figure 6.13.

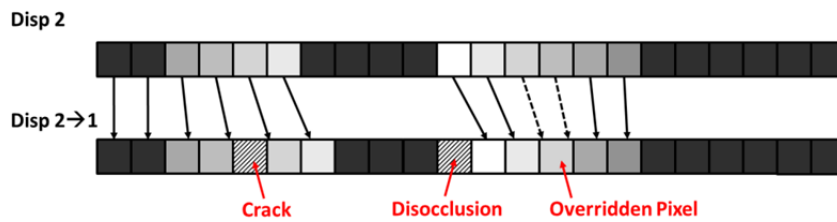


Figure 6.13. Schema of the naïve forward mapping DIBR process for disparity maps. Only integer pixel positions are allowed for the source and target pixels, hence artifacts similar to a nearest neighbor interpolation occur. If a target pixel, e.g. in the disparity maps $Disp_2 \rightarrow 1$ isn't hit by any source pixel, a crack or a disocclusion occurs. If a pixel is the target of two or more source pixels, the lower disparity is overridden. This is performed during a z-test which imitates the painter's algorithms, i.e. objects nearer to the observer occlude (and override) objects which are farther away.

$$\forall i \in [1, \text{size}(Disp_2)] : d_2 = Disp_{2(i)}, i_2 = i + \lfloor \beta_{21} \cdot d_2 \rfloor, d_{2 \rightarrow 1} = Disp_{2 \rightarrow 1}(i_2), \quad (6.3)$$

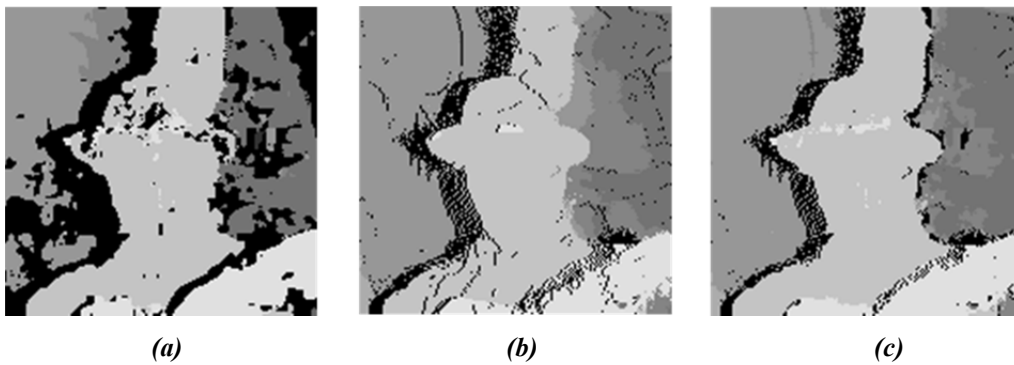
$$Disp_{2 \rightarrow 1}(i_2) = \begin{cases} d_2 & \text{if } d_2 > d_{2 \rightarrow 1} \\ d_{2 \rightarrow 1} & \text{else} \end{cases}$$

Figure 6.12 shows the result of the DIBR transfer of $Disp_2$ to the position of camera 1, resulting in a disparity map $Disp_2 \rightarrow 1$. As can be seen, small cracks appear at depth discontinuities. However, there is no need to spend special treatment as the disparity maps will be merged in the next step. If no

additional view was available for merging, sophisticated inpainting techniques as described in section 1.2.2.3 would be required.

6.3.8 Merging of the Satellite Disparity Maps

Once the disparity maps *Disp2* and *Disp3* have been rendered to the positions of the satellite cameras, the rendered disparity maps *D2→1* and *Disp3→4* are merged with the initial disparity maps *Disp12* and *Disp43*. The same algorithm is applied as described in sub-section 6.3.5. Again, inconsistent disparities are discarded while consistent disparities are averaged. As a result, much denser disparity maps for the satellite cameras are obtained. In Figure 6.14, a cut-out of the disparity maps before (b) and after (c) merging is shown. Most of the holes could be filled along the merging process.



*Figure 6.14. (a) Cut-out from the initial disparity map *Disp12*. (b) Cut-out from the rendered disparity map *Disp2→1*. (c) Cut-out from the result after merging.*

Nevertheless, for some pixels, no disparities are available. These disparities need to be filled in the subsequent bilateral filtering step. However, the quality of the bilateral filtering results increases when the input data is already dense and reliable.

6.3.9 Filtering the Satellite Disparity Maps

In a last step, the merged disparity maps need to be filtered using a cross-bilateral median filter [MüllerM10] in order to get pixel dense disparity maps. This is performed by applying the same algorithm as described in sub-section 6.3.6.

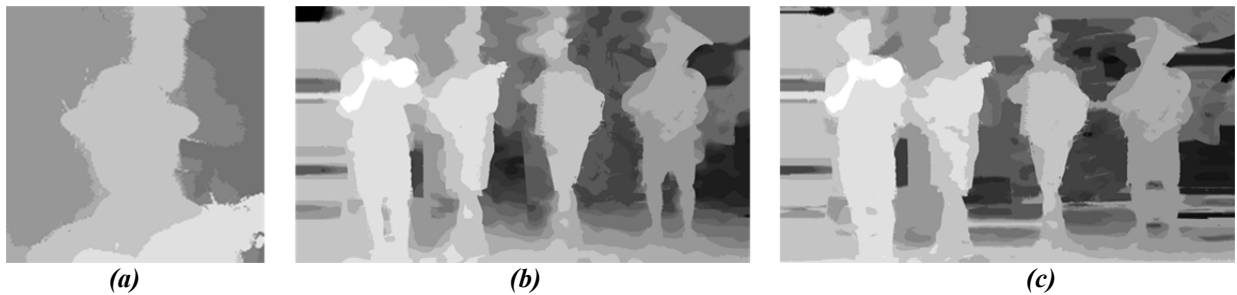


Figure 6.15. (a) Cut-out from filtered disparity map. (b) Filtered disparity map for camera 1. (c) If the previous merging step from section 6.3.8 is omitted, the filtering result contains remarkably more errors.

The result of the filtering of the disparity map is shown in Figure 6.15 (a, b). In comparison, to visualize the effect of the previous merging step, the merging step was omitted in Figure 6.15 (c). The

result of the filtering is shown which was applied to the disparity **Disp12** from Figure 6.12 (b) where no merging with the rendered disparity map **Disp2** \rightarrow **1** occurred. Apparently, the head of the leftmost musician now contains wrong disparities. In fact, the bilateral filter was not able to properly fill the corresponding hole in the disparity map from Figure 6.12 (b). The quality of the disparity in Figure 6.15 (b) is higher compared to the disparity map in Figure 6.15 (c).

6.4 Results

The resulting four pixel dense disparity maps which were generated using the stratified disparity estimation approach are shown in Figure 6.17. In the following, details on the evaluation of the data will be given. The proposed algorithm was applied to test-data which was captured in the research project MUSCADE [Muscade]. Four Sony HDC-P1 cameras with 2/3" sensors which have a native resolution of 1920x1080 pixels at a frame rate of 25 frames per second were used. Within the test shooting and experiments, a set of four Digiprime 10 mm lenses as fixed focal length lens, and four Fujinon HA 18x7.6 lenses as zoom lenses were used. For the test shooting the inter-axial distance between camera 2 and camera 3 was 60 mm (i.e. the narrow baseline), 240 mm between cameras 1 and 2 (left wide baseline) and 230 mm between cameras 3 and 4 (right wide baseline).

6.4.1 Offline MVD4 Generation

Four dense disparity maps were estimated using the proposed algorithm using test footage from the MUSCADE test sequence *Musicians 2*. The result can be seen in Figure 6.17. The disparity maps show a good alignment with the object boundaries. The density and accuracy of the disparity maps is similar among the four views although the corresponding stereo baselines differ considerably.



Figure 6.16. Test sequence *Musicians 2* shot during the 2nd MUSCADE [Muscade] test shooting. Original views of the cameras 1 to 4 from left to right. A small parallax is visible for the inner stereo pair, while a considerable parallax can be observed between the center and the satellite cameras.



Figure 6.17. Final disparity maps after all processing steps. In combination with the images from Figure 6.16, a MVD4 frame is composed. Please note that the disparity maps shown in this figure are shown using their original scaling. The disparity maps in section 6.3 were rescaled in order to expose artifacts.

However, no useful disparity information could be extracted from image regions which correspond to parts of the scene which were covered by one camera only. This was the case for the objects in the left part of **Disp1** (Figure 6.17, leftmost sub-image) and the right part of **Disp4** (Figure 6.17, rightmost sub-image) which are seen by a single camera only, hence a correct estimation of the corresponding

disparities was not possible. The corresponding disparity values are filled using a simple pixel repetition filter. By stitching the original video files from Figure 6.16 with the resulting disparity maps from Figure 6.17 a MVD4 frame is created. The goal of the generation of the MVD4 data is to generate views at virtual camera positions, and the quality of the generated views is therefore a relevant criterion. Consequently, a set of virtual views at nine different positions along the whole baseline was generated.

The resulting images are shown in Figure 6.18. The left-most and right-most views (views 1 and 9) coincide with the original views from camera 1 and camera 4 while all remaining views are synthesized using an approach described in [Kauff07]. The overall quality of the virtual views is convincing. However, the quality is slightly better near the original views (views 2, 4-6, and 8), compared to the views rendered in the center of the wide baselines, e.g. between camera 1 and camera 2 (view 3) and between the cameras 3 and 4 (view 6). To enhance the visibility of occurring rendering artifacts, a zoomed cut-out of the central musician in the scene is shown in Figure 6.19. The leftmost sub-image Figure 6.19 (a), corresponds to the original view from camera 2 and is shown for comparison purposes as it is not affected by DIBR related artifacts. The central sub-image Figure 6.19 (b) shows the interpolated view at a position which corresponds to the center of the narrow baseline stereo pair. Some artifacts are visible near the depth discontinuity between the musician's arm in the foreground and the blue background.

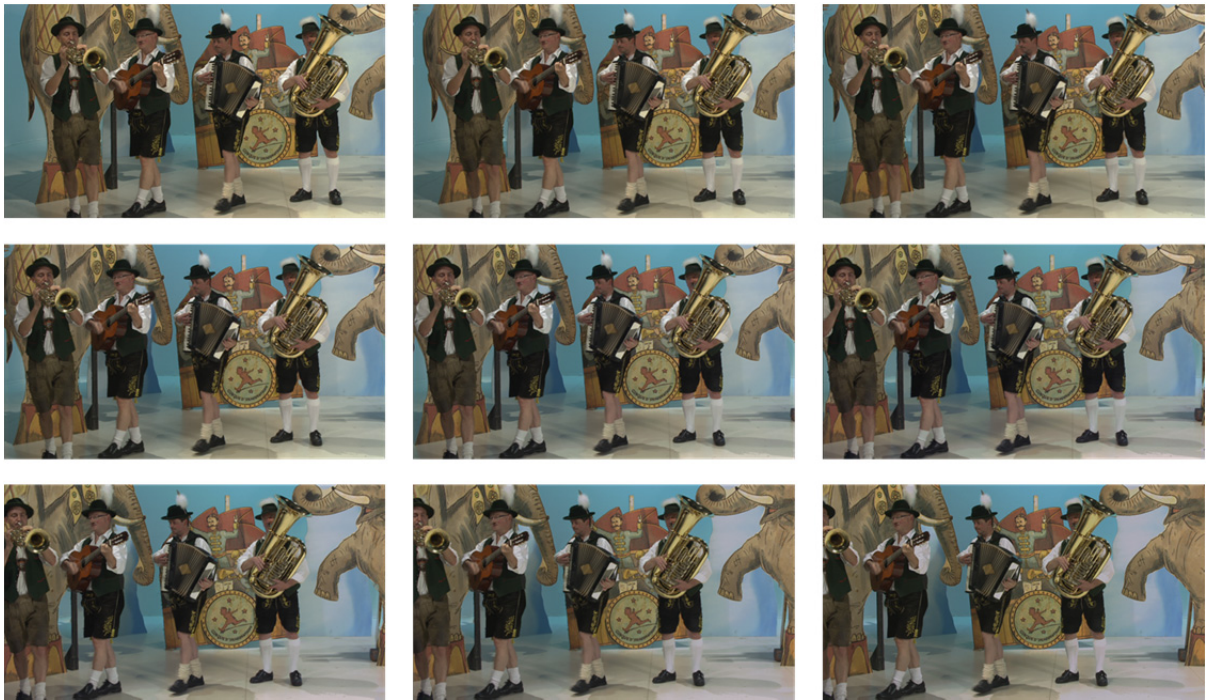


Figure 6.18. Virtual views 1-9 (counting from left to right and top to bottom) rendered along the wide baseline. The leftmost view in the top row coincides with the original camera 1, while the rightmost view in the bottom row coincides with the original camera 4. All remaining views are synthesized.

However, the parallax between this virtual view and the two nearest real views is small, which adds to the quality of the rendering. The rightmost sub-image Figure 6.19 (c) shows a virtual view at a

position centered between camera 3 and camera 4, i.e. between two cameras constituting a wide baseline. Apparently, the artifacts become stronger. The effect can be explained by inaccuracies in the disparity maps.



Figure 6.19. (a) Original view from camera 2. (b) Synthesized view between cameras 2 and 3 (narrow baseline) with small artifacts. (c) Synthesized view between cameras 3 and 4 (wide baseline), with stronger artifacts but still acceptable quality.

In fact, during the DIBR process, each pixel is shifted by an amount proportional to its normalized disparity and the virtual baseline between the real view and the virtual view. Noise in the disparity map is amplified by an amount proportional to this pixel shift. In consequence, an inaccuracy in the disparity map which might lead to a pixel positioning error of e.g. 1 pixel in the case of the narrow baseline, can lead to an offset of e.g. 4 pixels when rendering along the wide baseline. This emphasizes the importance of high quality disparity maps for DIBR in general and wide baseline DIBR in particular.

The proposed algorithm was implemented into a plug-in for the editing software VirtualDub and subsequently applied to other MUSCADE test sequences [Muscade]. Resulting MVD4 frames are shown in Figure 6.20 and Figure 6.23 which allow for a qualitative inspection of the disparity maps. The fully automatic and unsupervised disparity generation process was applied to all test sequences which were recorded using the multi-camera setup described above. The disparity maps show a similar quality along all four views except for image areas which are seen by a single camera only as for instance in the left-most and right-most part of Figure 6.23. The quality of the disparity maps appears to be comparable to the quality of the disparity map shown in Figure 6.17

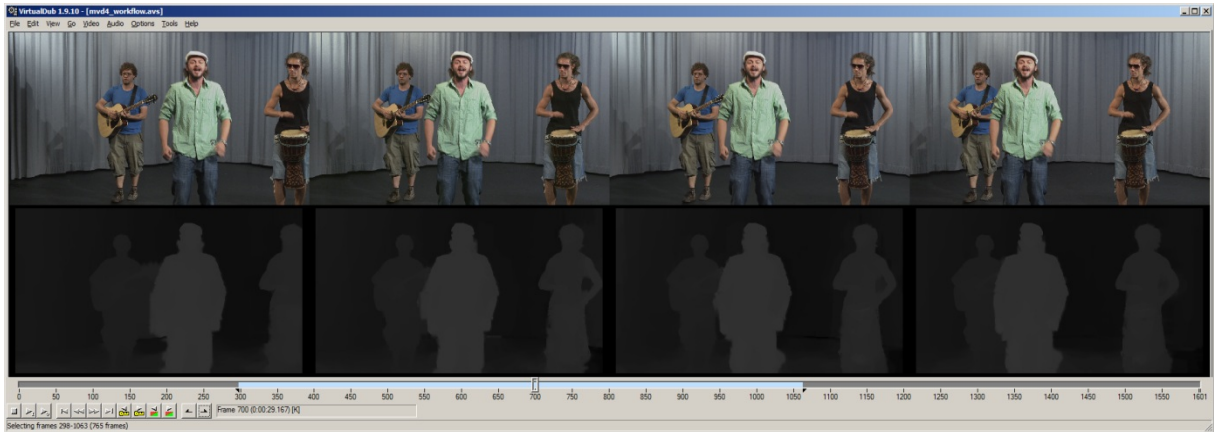


Figure 6.20. MVD4 Sequence Band 06 from the 1st MUSCADE test shooting generated inside a VirtualDub Plug-In implementing the mixed-baseline disparity estimation workflow proposed in chapter 6.



Figure 6.21. MVD4 Sequence BMX 04 from 2nd MUSCADE test shooting. This is a version using 720p50 resolution where the motion blur artifacts are significantly reduced. The fully automatic and unsupervised disparity generation process was applied.



Figure 6.22. MVD4 Sequence Musicians 2 from 2nd MUSCADE test shooting. The fully automatic and unsupervised disparity generation process was applied.



Figure 6.23. MVD4 Sequence Jungle 05 from 1st MUSCADE test shooting. The MVD4 generation for this content was particularly demanding due to the similarity of the colors of the foreground and background objects. In the image areas which are seen by only one camera (left-most and right-most part of the MVD4 picture), the disparities are unreliable. The fully automatic and unsupervised disparity generation process was applied.

6.4.2 Real-Time MVD4 Generation

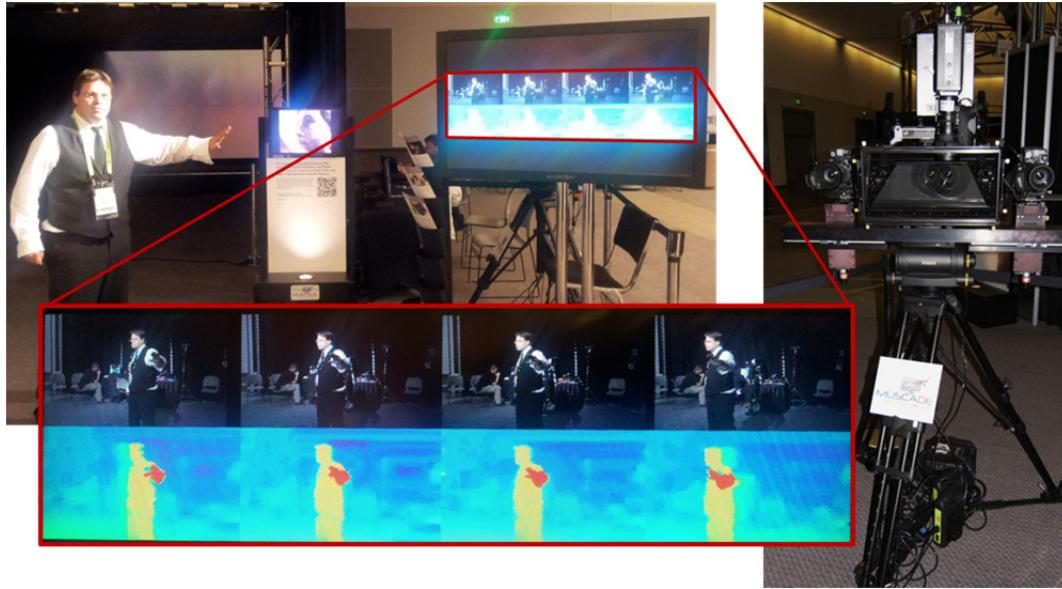


Figure 6.24. Real-Time MVD4 Generation as shown during Siggraph 2012 Emerging Technologies Campus [Kovacs12].

The real-time capabilities of the proposed MVD4 generation workflow were demonstrated in [Kovacs12] as shown in Figure 6.24. Beside the multi-camera rig, the system consisted of two PCs with an Intel Xeon X5550 Dual-Processor running the Windows 7 operating system with 64bit architecture performing the MVD4 related processing steps. A third PC of the same type pre-computed rectifying homographies using the algorithm described in section 0 for the four original camera views. The homographies were applied in real-time using the rectification engine described in section 5.3.4 as a pre-processing step. Figure 6.24 gives an impression of the real-time MVD4 content generation as shown on the Emerging Technologies Campus of Siggraph 2012 [Kovacs12]. As can be seen, a live captured four camera signal is processed into four corresponding disparity maps. The multi-camera rig which is mainly assembled from standard production equipment such as HD-TV cameras and a beam-splitter mounted on a conventional tripod and tripod-head is shown in the right part of Figure 6.24.

6.5 Conclusion

In this chapter, a stratified approach for the multi-camera disparity estimation using four cameras on a common baseline was proposed. A real-time capable demonstrator implementing the proposed algorithm was shown at Siggraph 2012 Emerging Technologies Campus [Kovacs12]. By combining disparities estimated from narrow and wide baselines, dense and consistent disparity maps with sub-pixel accuracy for all four cameras could be computed. The resulting quality of the disparity maps was assessed by creating virtual views along the whole baseline using depth-image-based rendering (DIBR). The high quality of the resulting renderings was demonstrated using different test sequences created within the MUSCADE project [Muscade]. However, the accuracy of the disparity maps has shown to be an important factor for the resulting quality. A depth resolution of 8 bit per pixel is clearly

not enough when coping with wide baselines. Good DIBR results could be achieved using a depth resolution of 10 bits per pixel. Remaining rendering artifacts can be neglected if the content is not shown on a display with Full HD resolution per view, which is the case for today's auto-stereoscopic displays. In contrast, many glasses-based stereoscopic 3D-TVs offer a Full-HD resolution per view. Consequently, the artifacts caused by DIBR become exposed when shown on these displays and impair the visual quality. Hence, a native stereoscopic image pair e.g. captured using a beam-splitter is required for these displays.

7 Conclusion and Outlook

7.1 Summary

The aim of this thesis was the development of image processing algorithms which are suitable to facilitate the creation of high quality multi-camera content, e.g. for stereoscopic 3D and multi-view devices. As carried out in section 1.1.1 (Problem Statement and Relevance), this is of great importance for the 3D movie industry, because high quality stereoscopic 3D content is needed for a high quality of experience while wrongly produced 3D content can lead to eye strain and visual fatigue. Against this background, a set research questions were derived in section 1.1.2, asking the question which image processing algorithms in particular might be suitable to facilitate the content creation process, e.g. image rectification, feature detection and matching, algorithms for the assisted 3D production or depth based content creation. To allow a structured treatment of the research questions, the state-of-the-art of the current 3D production workflow including different production schemes such as native 3D production, 2D to 3D conversion and today's depth-based content production approaches were described in section 1.2. The introductory chapter 1 closes with a general overview of the dissertation in section 1.3. An overview of publications related to the publication was given (1.3.1) as well as a summary of the main contributions and novelties (section 1.3.2). The structure of the dissertation was presented in section 1.3.3.

In chapter 2, the theoretical background for the chapters 3 to 6 was presented giving details about the geometry of stereoscopic 3D content reproduction (section 2.1), the human visual system (HVS) and depth perception in section 2.2. Production rules for the acquisition of 3D content which have to be respected by stereographers were presented in section 2.3. In section 2.4, fundamental concepts of the projective geometry along with basic concepts of the stereo and multi-camera rectification were described. In section 2.5, basic concepts of feature detection were presented, before completing the chapter with section 2.6 where fundamental concepts of disparity estimation techniques were presented.

In the introduction of chapter 3 the concept of the Taylor expansion was applied to projective entities (section 3.1). Subsequently, this concept was applied to the linearized estimation of the fundamental matrix (section 3.2) and the estimation of the trifocal tensor (section 3.3). The linearized projective entities were then used to derive rectification algorithms specialized for the case of nearly rectified stereo cameras (section 3.2) and linear cameras arrays (section 3.3) along with a quantitative evaluation of the performance of the proposed algorithms.

In chapter 4 a feature descriptor called semantic kernels binarized (SKB) was presented. After giving details on the basic properties of the descriptor (section 4.2), the several steps of the description process, i.e. the definition of the support region, the sampling of the support region (section 4.3) and

the folding with a set of binary kernels (section 4.4) were described. Matching strategies which take advantage of the binary feature vector of the descriptor were presented in section 4.5. A comparison with state-of-the-art feature point descriptors was performed in section 4.6 along with a detailed quantitative evaluation.

In chapter 5, algorithms for the assisted production of stereoscopic 3D content were described which include temporal consistent camera pose estimation, rectification and disparity histogram analysis. The algorithms are based on the approaches for the linearized estimation of the fundamental matrix from chapter 3 and the feature descriptor SKB from chapter 4. An overview of the components of the assistance system was given in section 5.3. The basic algorithms were subsequently used to derive the near and the far clipping plane of a scene in order to derive optimal stereoscopic settings for the convergence plane and the inter-axial distance (section 5.4). A detailed comparison of the updated production workflow with legacy production workflows was performed in section 5.5. The chapter 5 was concluded in section 5.6.

In chapter 6, a stratified algorithm for the mixed baseline disparity estimation was presented which was applied to a multi-camera setup described in section 6.2. The setup included a multi-camera rectification based on the trifocal tensor estimated using the approach from chapter 3, and feature points matched using the approach from chapter 4. In section 6.3, the stratified approach for the multi-camera disparity estimation was presented which combined disparity estimation from a narrow and a wide baseline. The narrow baseline allowed the estimation of dense disparity maps while the wide baseline was suitable to estimate sparser disparity maps but with higher depth accuracy. Results of the approach were presented in section 6.4.

7.2 Main Contributions

In chapter 1, an overview of the field of stereoscopic 3D production is given, along with an overview of the relation between different image processing techniques and 3D content production. In chapter 2, theoretical foundations of the 3D production process based on insights about the human visual system were given. In addition, concepts of underlying image processing techniques such as projective geometry, stereo- and multi-rectification, feature detection, description and matching, and disparity estimation were explained. The chapters 3 to 6 are the chapters of the dissertation in which the algorithmic contributions were presented and evaluated. In chapter 3, stereo and multi-camera rectification algorithms are presented. In chapter 4, a feature descriptor (SKB) was presented. An assistance system for the simplified production of stereoscopic 3D content is presented in chapter 5. In chapter 6, a mixed-baseline disparity estimation algorithm was presented.

An overview of the main novelties and contributions was given in the sub-sections 1.3.2.1 to 1.3.2.5. The thesis contains the following contributions:

- An overview of the state-of-the-art of stereoscopic 3D production;
- An overview of the theoretical background of adjacent technologies such as 3D reproduction, projective geometry, feature point matching and disparity estimation;
- An algorithm for the estimation of a linearized fundamental matrix along with a stereo rectification and pose estimation algorithm;
- An algorithm for the estimation of the linearized trifocal tensor and a multi-camera rectification algorithm;
- A feature point descriptor called semantic kernels binarized (SKB);
- A set of algorithms for the simplified 3D production, e.g. algorithms for the automated and time-consistent adaptation of the inter-axial distance, convergence plane, and correction of the stereo camera alignment;
- An algorithm for the disparity estimation using a mixed narrow and wide stereo baseline.

A list of publications related to this thesis was given in section 1.3.1. In the following, the algorithmic contributions are summarized.

7.2.1 Linearized Fundamental Matrix

A stereo camera rectification method which performs reliably and which uses only point correspondences which can directly be extracted from the stereo image pairs, was presented in section 3.2. The rectification algorithm was applied to stereoscopic 3D sequences as described in chapter 5, i.e. the rectification method needed to minimize any possible distortion in order to generate visually pleasant stereo pairs. In addition, the convergence plane should not be changed.

The concept of the linearized computation of the fundamental matrix as described in chapter 3 was first published in [Zilly10a]. It has since then attracted the attention of the scientific community as can be seen in the references [Heinzle11] and [Georgiev13] whereby the latter performs an in-depth comparison of the algorithm described in this thesis with the contribution from Georgiev et al.

7.2.2 Linearized Trifocal Tensor

In section 3.3 a new method for a robust estimation of the trifocal tensor specialized for linear camera arrays and subsequent rectifying homography computation based on feature point triplets was proposed. It was thereby assumed that the geometric configuration is not far from the rectified state and that consequently a linearization is possible, which for instance was given in the setup described in chapter 6. The algorithm achieves vertical alignment and horizontal alignment, i.e. proportional horizontal disparities after rectification. It furthermore is able to estimate the ratio of the camera baselines which do not need to be equidistant. It is based on feature point triplets and suitable for uncalibrated cameras. The proposed estimation method for the computation of the linearized trifocal tensor was first published in [Zilly12c].

7.2.3 Feature Point Descriptor SKB

A binarized descriptor which has a low memory usage and good matching performance was proposed. The descriptor is composed of binarized responses resulting from a set of folding operations applied to the normalized support region. A main property of the SKB is a lower computational load and complexity. Its fast run-time enables near real-time updates of stereo rectification parameters. Details of the feature descriptor SKB were presented in chapter 4. The descriptor SKB was first proposed in [Zilly11c]. It has since then attracted the interest of the research community, e.g. it was used by Stefanoski et al. [Stefanoski13] as part of the framework of the Image Domain Warping algorithm, an approach which was submitted to MPEG resulting in one of the four best proposals in the multi-view-autostereoscopic display test scenario. Furthermore, the SKB was implemented as ASIC core by Schaffner et al. in [Schaffner13] which is able to process 25000 Interest Points at 720p resolution in real-time.

7.2.4 Algorithms for the Simplified 3D Production

A set of new algorithms for the temporal consistent estimation of the 3D camera geometry, such as the generation of a disparity histogram for the derivation of the near and far clipping plane, were presented in chapter 5. The algorithms combine and make use of the rectification algorithm from chapter 3 and the feature descriptor SKB from chapter 4. In combination with a PC system with graphical user-interface, the algorithms are the core of a camera assistance system which supports the stereographer using comfort functions such as the automatic derivation of the convergence plane and the inter-axial distance. Concepts and demonstrators of the assistance system were first presented in 2009 at the NAB Show in Las Vegas and presented in [Zilly09], [Zilly10b] and [Zilly11b] to the international scientific community. Please note that the majority of the algorithms, derivations and explanations performed in chapter 5 have not been previously published or never been described in this level of detail before. The ideas presented in the papers and this thesis however, had significant influence on the literature on one hand and commercial products on the other hand. Several high ranked publications such as [Heinzle11], [Greisen11], [Oskam11], [Celikcan13] and [Templin14] are good examples for the influence of these papers. The influence on the 3D industry was also important, as for instance the first real-time transmission of a 3D live concert was performed using the system in 2010 [Wagner10] while the assistance system has since then also become part of a commercial product line [Schmidt11]²⁰. It is furthermore mentioned in the rather industry-inclined publication from Mendiburu et al. ([Mendiburu12], page 89).

7.2.5 Mixed Baseline Multi-View Video plus Depth Generation

A multi-camera disparity estimation algorithm dedicated for the use of mixed stereo baselines was presented which is suitable for real-time execution. The setup is based on a four camera rig involving a

²⁰ The stereoscopic analyser STAN is part of the product lines VENICE and CLIPSTER from the Rhode&Schwarz Company DVS, <http://www.dvs.de/>

central narrow baseline, with two cameras mounted on a standard beam-splitter rig known from stereoscopic 3D productions, and a wide baseline comprising of two satellite cameras mounted outside the mirror box. The algorithm was pre-published in [Zilly12b] and [Zilly14]. It has since then attracted the attention of the research community, resulting in citations of high ranked papers such as [Chapiro14] and [Baek14]. Moreover, a demonstrator of the proposed algorithm was shown at the Emerging Technologies Campus at Siggraph 2012 [Kovacs12].

7.3 Discussion & Outlook

It was shown that digital image processing can greatly improve the quality and efficiency of 3D signal processing. Nevertheless, a question which remains open is, if the 3D technology can be established permanently. It is fair enough to say, that from a today's point of view, it is too early for a final answer. An often discussed fact is that there were earlier tries to establish 3D in the market (50s, 70s, etc.) and that these attempts mostly failed [Zilly11b]. Beside all technical improvements of the last years, one cannot ignore, that huge challenges remain regarding a successful introduction of 3D-TV. In this context, one can argue, what is necessary or missing for a successful introduction of 3D-TV? And regarding 3D cinema, is this really a sustainable development? One can argue that giving the fact that 3D came back many times shows us, that stereoscopic 3D has the potential to improve the overall quality of experience in entertainment industry. Moreover, the 3D technology takes advantage of improvements in the field of 2D, for instance, higher resolutions (e.g. 4k) and high frame rates. It might be successful as soon as the minimum requirements are met in terms of quality and production costs. It was described in chapter 1 that 3D requires a minimum quality, because bad 3D hurts. This can be comparable to a minimum framerate for a video to be recognized as *moving picture*. In that context, the proposed assistance system from chapter 5 promises to simplify a high quality stereoscopic 3D production workflow. The underlying algorithms which were mainly described in chapter 3 and chapter 4 have shown to be suitable image processing algorithms.

Furthermore, the diversity of displays on which 3D content shall be reproduced is growing. Beside 3D cinemas and 3D-TVs for the living room, autostereoscopic tablets and head-mounted displays could be used in the future. In that context, the proposed workflow from chapter 6 promises to facilitate the content creation and adaptation. It enables a depth-based rendering approach which allows for generating virtual stereo baselines which can be adapted to the 3D device on one hand, and user preferences on the other hand.

8 Appendix

8.1 Stereoscopic Test Productions

In this section, example of field tests and test productions of the 3D assistance system called STAN are presented²¹. The assistance system was used within several test productions under real conditions. Thereby, it was used to support the 3D production in the following two scenarios:

- On set as assistance for the mechanical alignment of the cameras and a proper choice of the inter-axial and convergence distance.
- As stereo image processor with real-time correction for live productions where the STAN was operated by stereographers inside an OB van.

Certainly, one of the highlights was the live transmission of the pop concert of Germany's Hip-Hop-Band "Die Fantatischen Vier" which was broadcasted live from Halle/Saale into over 90 cinemas in 5 European countries. In cooperation with ARTE, STAN has also been used during a live-on-tape production at the New Pop Festival 2010 in Baden-Baden featuring the British pop band "Marina and the Diamonds". Other important productions were the first German 3D short movie "Topper gibt nicht auf" produced by the Film and Television University "Konrad Wolf" in Potsdam Babelsberg and the 3D recording of an orchestral rehearsal of the Berlin Philharmoniker with Sir Simon Rattle. In March 2010, the live broadcast abilities have been shown at Cebit 2010 in Hanover, Germany.

In the following, a more detailed description of the above-mentioned 3D productions is given.

8.1.1 Cebit 2010

A first test for a live broadcast under controlled conditions was the live 3D transmission showcase of the Fraunhofer booth at Cebit 2010 in Hanover, Germany. This setup was conducted jointly with Fraunhofer IIS and KUK Filmproduktion. Two microHD cameras from Fraunhofer IIS were mounted on a side-by-side rig. STAN was used to roughly align the two cameras with manual interaction of a stereographer. However, the side-by-side rig had only limited degrees of freedom for the calibration – a property which is typical for small sized stereo rigs. Remaining roll and tilt errors could not be equalized completely by mechanical adjustments only. In addition, the focal lengths of the two fixed focal length C-mount lenses differed by around 1-2%.

Therefore, as it was not possible to achieve a perfect alignment mechanically, a geometric correction was applied electronically in real-time by a PC running the STAN. Furthermore, as the two cameras were mounted using a parallel setup, a suitable convergence plane had to be set electronically.

²¹ The description of the test productions including pictures in this appendix have been previously published in [Zilly11a].

Subsequently, the corrected HD-SDI stream was encoded using an Ultra-Low-Delay H.264 encoder and transmitted via TCP/IP to a decoding unit which rendered the stereoscopic video stream on a set of 3D-TV displays.



Figure 8.1. Josef Kluger from KUK Film Production presenting the 3D transmission showcase of the Fraunhofer booth at Cebit 2010.

Figure 8.1 shows Josef Kluger, CEO of KUK Filmproduktion explaining the system to the audience during a press conference with Prof. Dr.-Ing. Hans-Jörg Bullinger, former President of the Fraunhofer-Gesellschaft.

8.1.2 Berliner Philharmoniker

Another 3D production with the STAN was the recording of an orchestral rehearsal of the Berlin Philharmonic Orchestra (BPO) with Sir Simon Rattle.



Figure 8.2. 3D Production showing an orchestral rehearsal of Berlin Philharmonic Orchestra.

The STAN system has been used for the calibration of the stereo rigs. The calibration process was improved in accuracy while the time duration could be reduced. Figure 8.2 shows the scenery of the production set.

8.1.3 Fantastische Vier

A major event in 2010 was the 3D live transmission of a concert of the German Hip-Hop band “Die Fantastischen Vier” in more than 90 cinemas in five European countries. In total, five stereo rigs were used for the live transmission, including four mirror rigs and one side-by-side rig. The calibration of the mirror rigs was conducted with assistance of the STAN. In addition, all remaining vertical disparities were corrected by electronic image rectification. Five dual HD-SDI streams were corrected in real-time using the STAN running on DVS Venice video server inside an OB van. Figure 8.3 shows a screenshots of the stereographers’ operation desk.



Figure 8.3. *STAN inside the OB van which was used to broadcast the 3D live concert.*

The task of the stereographers from KUK Filmproduktion was to shift convergence plane and to adapt the inter-axial distance dynamically while shooting. The former was changed directly by an electronic sensor shift via STAN, the latter was changed manually on the rig. A wireless intercom was used to ensure communication between the personnel inside the OB van and the stereographers on stage.

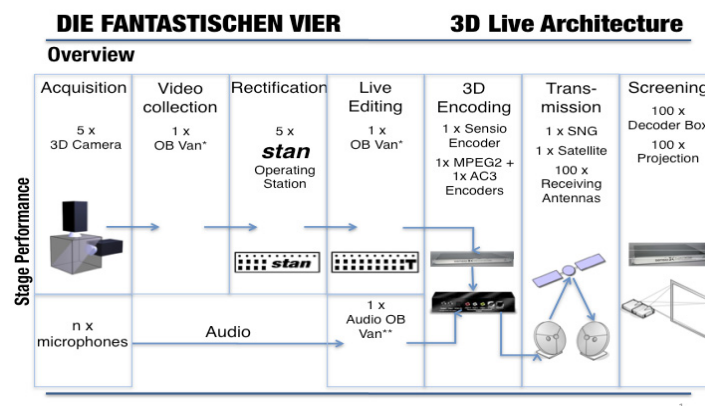


Figure 8.4. *Architecture and flowchart of 3D Live Production.*

After rectification, the signal was ingested into a vision mixer for live editing and switching between the different stereo streams. The program stream was then encoded and transmitted using a SNG to the broadcast satellite. The signal was received by over 90 cinemas. Figure 8.4 gives an overview of the overall architecture and signal flow.

8.1.4 Marina and the Diamonds

In cooperation with ARTE the STAN has also been used during a live-on-tape production at the SWR3 New Pop Festival 2010 in Baden-Baden featuring the British pop band “Marina and the Diamonds”. Five stereo rigs were used for live production and recording. The image rectification was performed using STAN on a DVS video server inside an OB van.



Figure 8.5. Inside the OB van the signal of the five stereo cameras was analysed. The STAN was used during the calibration process of the stereo rigs and for the real-time rectification of the stereo streams.

Figure 8.5 shows the stereographers during the calibration process which was guided from the STAN operators inside the OB van.



Figure 8.6. 3D transmission of the performance of “Marina and the Diamonds” during the New Pop Festival.

Figure 8.6 shows impressions of the 3D live transmission. One beam-splitter rig was mounted on a crane. The cameras were equipped with zoom lenses and different focal lengths were used during the transmission. This raised the need for an adaptive image rectification. During this event, new parameters for the geometrical correction had to be estimated after changing the focal length. This was conducted while the camera pair was in the off. In the meantime, a robust dynamic auto-correction feature had been implemented with respect to the feedback and the experience gathered within this test production.

9 Glossary

9.1 Technical Terms

Blobness	Measure used to identify possible interest point when using Blob detectors such as SIFT or SURF
Convergence Puller	Person on a 3D film-set which controls the convergence distance
Depth Budget	Upper limit for the amount of depth for a 3D image pair allowing a comfortable 3D reproduction
Diplopia	Double vision
Ghosting	Partial overlay of the right or left stereo image due to insufficient channel separation
Keystone	Geometric distortion due to convergent stereo camera setup
Principal Point	Intersection of the optical axis with the image plane
Scale Space	Set of 2D images derived from the same original image, folded using different kernels, e.g. using Gaussian kernels with increasing standard deviation σ .
Screen Space	Objects in the screen space are perceived behind the screen
Stereographer	Person in charge of ensuring that all technical and creative stereoscopic parameters are properly set
Support Region	Entity of pixels around an interest point used to compute the descriptor
Viewer Space	Objects in the viewer space are perceived in front of the screen

9.2 Abbreviations and Acronyms

720p50	Video raster with 1280x720 pixels and 50 progressive frame per second
CVR	Comfortable Viewing Range
DIBR	Depth Image Based Rendering
DoG	Difference of Gaussian
FIFO	First In First Out
HD-SDI	High Definition Serial Digital Interface
HIT	Horizontal Image Translation
HVS	Human visual system
LoG	Laplacian of Gaussian
MVD	Multi-view Video plus Depth
MVD4	MVD with 4 video and 4 depth streams
SAD	Sum of Absolute Differences
STAN	Stereoscopic Analyzer

9.3 Latin and Mathematical Symbols

A	Constraint matrix used to set up a linear system of equations
b	Constraint vector used to set up a linear system of equations
B	(Stereo-) Baseline, inter-axial distance
C	Camera center
d	Disparity (can be measured in pixels or similar unit)
\tilde{d}	Relative disparity (i.e. disparity divided by sensor width)
E	Essential matrix
F	Fundamental matrix
H	Homography
I	Identity matrix
K	Intrinsic matrix
L_{AC}	Parallax-limit induced by the accommodation-convergence conflict
M	Coordinates of a world point in 3D space
m	2D projection of a 3D world point in camera coordinates
P	Projection matrix
\mathcal{P}	Screen Parallax (can be measured in cm or similar unit)
$\tilde{\mathcal{P}}$	Relative screen parallax, i.e. screen parallax divided by screen width
R	Rotation matrix
t_{eye}	Inter-ocular distance (eye distance)
\mathcal{T}_i^{jk}	Trifocal tensor
u	Horizontal pixel coordinate, possibly in the left or first camera
u'	Horizontal pixel coordinate, possibly in the right or second camera
u''	Horizontal pixel coordinate in the third camera
v	Vertical pixel coordinate, possibly in the left or first camera
v'	Vertical pixel coordinate, possibly in the right or second camera
v''	Vertical pixel coordinate in the third camera
x	Result vector
Z_D	Distance between viewer and screen
Z_v	Distance between viewer and perceived object

Bibliography

10.1 Publications by the Author

- [Feldmann08] I. Feldmann, M. Müller, F. Zilly, R. Tanger, K. Müller, A. Smolic, P. Kauff, and T. Wiegand. HHI test material for 3D video. ISO/IEC JTC1/SC29/WG11, M15413, Archamps, France, April 2008.
- [Hubert13] H. Hubert, B. Stabernack, and F. Zilly. Architecture of a Low Latency Image Rectification Engine for Stereoscopic 3-D HDTV Processing. *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 23, issue 5, pp. 813-822, 2013.
- [Kauff10] P. Kauff, F. Zilly, R. Schäfer, J. Kluger, and N. Malchow. Stereoscopic Analyzer (STAN). *Fachzeitschrift für Fernsehen, Film und Elektronische Medien (FKT)*, 4/2010, pp. 178-184, 2010.
- [Kauff12] P. Kauff, K. Müller, R. Schäfer, and F. Zilly. Dreidimensional fernsehen. *Physik in unserer Zeit*, vol. 43, issue 3, pp. 116-223, May 2012.
- [Kovacs12] P. T. Kovacs and F. Zilly. 3D capturing using multi-camera rigs, real-time depth estimation and depth-based content creation for multi-view and light-field auto-stereoscopic displays. *ACM SIGGRAPH 2012 Emerging Technologies*, no. 1, Los Angeles, USA, Aug. 2012.
- [MüllerM10] M. Müller, F. Zilly, and P. Kauff. Adaptive cross-trilateral depth map filtering. In *3DTV-Conference (3DTV-CON)*, Tampere, Finland, June 2010.
- [MüllerM11] M. Müller, F. Zilly, C. Riechert, and P. Kauff. Spatio-temporal consistent depth maps from multi-view video. In *3DTV-Conference (3DTV-CON)*, Antalya, Turkey, May 2011.
- [Riechert12a] C. Riechert, F. Zilly, P. Kauff, J. Güther, and R. Schäfer. Fully automatic stereo-to-multiview conversion in autostereoscopic displays. *The Best of IBC and IET*, vol. 4, pp. 8-14, *Best Paper*, Sept. 2012.
- [Riechert12b] C. Riechert, F. Zilly, M. Müller, and P. Kauff. Advanced Interpolation Filters for Depth Image Based Rendering. In *3DTV-Conference (3DTV-CON)*, Zurich, Switzerland, Oct. 2012.
- [Riechert12c] C. Riechert, F. Zilly, M. Müller, and P. Kauff. Real-Time Disparity Estimation Using Line-Wise Hybrid Recursive Matching and Cross-Bilateral Median Up-Sampling. In *International Conference on Pattern Recognition (ICPR)*, Tsukuba, Japan, Nov. 2012.
- [Zilly09] F. Zilly, P. Eisert, and P. Kauff. Real-Time Analysis and Correction of Stereoscopic HDTV Sequences. In *Conference on Visual Media Production (CVMP), Short Paper*, London, UK, Nov. 2009.
- [Zilly10a] F. Zilly, M. Müller, P. Eisert, and P. Kauff. Joint Estimation of Epipolar Geometry and Rectification Parameters using Point Correspondences for Stereoscopic TV Sequences. In *5th International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, Paris, France, May 2010.
- [Zilly10b] F. Zilly, M. Müller, and P. Kauff. The Stereoscopic Analyzer – An Image-Based Assistance Tool for Stereo Shooting and 3D Production. In *International Conference on Image Processing (ICIP), Special Session on Image Processing for 3D Cinema Production, Invited Paper*, Hong Kong, Sept. 2010.
- [Zilly11a] F. Zilly, M. Müller, P. Kauff, and R. Schäfer. STAN — An assistance system for 3D productions: From bad stereo to good stereo. In *14th ITG Conference on Electronic Media Technology (CEMT)*, Dortmund, Germany, March 2011.
- [Zilly11b] F. Zilly, J. Kluger, and P. Kauff. Production Rules of 3D Stereo Acquisition. *Proc. of the IEEE (PIEEE), Special Issue on 3D Media and Displays*, vol. 99, issue 4, pp. 590-606, *Invited Paper*, April 2011.
- [Zilly11c] F. Zilly, C. Riechert, P. Eisert, and P. Kauff. Semantic Kernels Binarized - A Feature Descriptor for Fast and Robust Matching. In *Conference on Visual Media Production (CVMP)*, London, UK, Nov. 2011.
- [Zilly12a] F. Zilly, N. M. Gutberlet, C. Riechert, R. Tanger, and P. Kauff. Depth Based Content Creation Targeting Stereoscopic and Auto-Stereoscopic Displays. In *International Broadcasting Convention (IBC)*, Amsterdam, NL, Sept. 2012.
- [Zilly12b] F. Zilly, C. Riechert, M. Müller, and P. Kauff. Generation of multi-view video plus depth content using mixed narrow and wide baseline setup. In *3DTV-Conference (3DTV-CON)*, Zurich, Switzerland, Oct. 2012.
- [Zilly12c] F. Zilly, C. Riechert, M. Müller, W. Waizenegger, T. Sikora, and P. Kauff. Multi-Camera Rectification using Linearized Trifocal Tensor. In *International Conference on Pattern Recognition (ICPR)*, pp. 2727-2731, Tsukuba, Japan, Nov. 2012.
- [Zilly13] F. Zilly, M. Müller, and P. Kauff. Generic Content Creation for 3D Displays. In *3D-TV System with Depth-Image-Based Rendering*, Springer New York. Ce Zhu, Yin Zhao, Lu Yu, Masayuki Tanimoto (Editors), pp. 39-68, Jan. 2013.
- [Zilly14] F. Zilly, C. Riechert, M. Müller, P. Eisert, T. Sikora, P. Kauff. Real-time generation of multi-view video plus depth content using mixed narrow and wide baseline. *Journal of Visual Communication and Image Representation, Special Issue on 3D Video Processing*, vol. 25, issue 4, pp. 632-648, 2014.

10.2 Other Publications

- [3ality] 3ality Digital. Stereoscopic Image Processor SIP2100. <http://3alitydigital.com>.
- [3D4YOU] 3D4YOU, Content Generation and Delivery for 3D Television. *Seventh Framework Theme ICT-2007.1.5 Networked Media*. Grant Agreement no. 215075. http://cordis.europa.eu/project/rcn/85534_en.html
- [Alais05] David Alais and Randolph Blake (Editors). *Binocular Rivalry*. Edited by, MIT Press, Cambridge, MA, ISBN: 0-262-01212-X, 2005.
- [An04] L. An, Y. Jia, J. Wang, X. Zhang, and M. Li. An efficient rectification method for trinocular stereovision. In *International Conference on Pattern Recognition (ICPR)*, vol. 4, pp. 56–59, Washington, DC, USA, 2004.
- [Atzpadin04] N. Atzpadin, P. Kauff, and O. Schreer. Stereo Analysis by Hybrid Recursive Matching for Real-Time Immersive Video Stereo Analysis by Hybrid Recursive Matching for Real-Time Immersive Video Conferencing. *IEEE Trans. on Circuits and Systems for Video Technology, Spec. Issue on Immersive Telecomm.*, vol. 14, issue 3, pp. 321-334, Jan. 2004.
- [Ayache88] N. Ayache and C. Hansen. Rectification of images for binocular and trinocular stereovision. In *International Conference on Pattern Recognition (ICPR)*, vol. 1, pp. 11–16, Rome, Italy, Nov. 1988.
- [Baek14] S.-H. Baek and M. H. Kim. Stereo Fusion using a Refractive Medium on a Binocular Base. In *12th Asian Conference on Computer Vision (ACCV'14)*, Singapore, Nov. 2014.
- [Baik07] Y. K. Baik, J. H. Choi, and K. M. Lee. An Efficient Trinocular Rectification Method for Stereo Vision. *Proc. Frontiers of Computer Vision (FCV)*, Jan. 2007.
- [Balogh07] T. Balogh, P. Kovacs, and A. Barsi. Holovizio 3D display system. In *3DTV-Conference (3DTV-CON)*, Kos Island, Greece, May 2007.
- [Bartczak11] B. Bartczak, P. Vandewalle, O. Grau, G. Briand, J. Fournier, P. Kerbiriou, M. Murdoch, M. Müller, R. Goris, R. Koch, and R. van der Vleuten. Display-Independent 3D-TV Production and Delivery Using the Layered Depth Video Format. *IEEE Trans. on Broadcasting*, vol. 57, issue 2, part 2, pp. 477-490, June 2011.
- [Bay08] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool. SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding (CVIU)*, vol. 110, no. 3, pp. 346-359, 2008.
- [Beldie91] I. P. Beldie, B. Kost. Luminance asymmetry in stereo TV images. *Proc. SPIE, Stereoscopic Displays and Applications II*, vol. 1457, pp. 242-247, Aug. 1991.
- [Belhumeur96] P. N. Belhumeur. A Bayesian approach to binocular stereopsis. *International Journal of Computer Vision (IJCV)*, vol. 19, issue 3, pp. 237–260, Aug. 1996.
- [Bergmann93] Ludwig Bergmann, Clemens Schaefer, Heinz Niedrig. *Lehrbuch der Experimentalphysik, Bd.3, Optik, 9. Auflage*. Walter de Gruyter, Berlin, 1993.
- [Binocle] Binocle. DispartiyTagger & DispartiyKiller. <http://www.binocle.com>
- [Birchfield98a] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 20, issue 4, pp. 401-406, Apr. 1998.
- [Birchfield98b] S. Birchfield and C. Tomasi. Depth Discontinuities by Pixel-to-Pixel Stereo. In *International Conference on Computer Vision (ICCV)*, pp. 1073-1080. Bombay, India, Jan. 1998.
- [Bleyer07] M. Bleyer and M. Gelautz. Graph-cut-based stereo matching using image segmentation with symmetrical treatment of occlusions. *Signal Processing: Image Communication*, vol. 22, issue 2, pp. 127-143, Feb. 2007.
- [Bolles87] R. C. Bolles, H. H. Baker, and D. H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision (IJCV)*, vol. 1, issue 1, pp. 7-55, 1987.
- [Boutarel10] F. Boutarel and V. Nozick. Epipolar rectification for autostereoscopic camera setup. In *6th Europe-Asia Congress on Mechatronics*, pp. 133-136, Yokohama, Japan, Nov. 2010.
- [Boykov01] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 23, issue 11, pp. 1222-1239, Nov. 2001.
- [Brandt10] J. Brandt. Transform coding for fast approximate nearest neighbor search in high dimensions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1815-1822, 2010.
- [Brewster56] D. Brewster, *The stereoscope: it's history, theory and construction*, London, 1856.

- [Bronstein95] I. N. Bronstein, K. A. Semendjajew, G. Musiol, and H. Mühlig, *Taschenbuch der Mathematik*, Harry Deutsch, 1995.
- [Brown02] M. Brown and D. Lowe. Invariant features from interest point groups. In *British Machine Vision Conference (BMVC)*, Cardiff, UK, Sept. 2002.
- [Brown03] M.Z. Brown, D. Burschka, and G.D. Hager. Advances in Computational Stereo. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 25, issue 8, pp. 993-1008, Aug. 2003.
- [Buchs11] J. Buchs, S. Heimbecher, and A. Reitano. Ein Jahr „Sky 3D“ – Produktion, Sendeabwicklung und Übertragung in der dritten Dimension. *Fachzeitschrift für Fernsehen, Film und Elektronische Medien (FKT)*, 8-9/2011, pp. 430-434, 2011.
- [Calonder10] M. Calonder, V. Lepetit, and P. Fua. BRIEF: Binary Robust Independent Elementary Features. In *European Conference on Computer Vision (ECCV), Part IV*, pp. 778-792, Heraklion, Crete, Greece, Sept. 2010.
- [Campbell08] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using Multiple Hypotheses to Improve Depth-Maps for Multi-View Stereo. In *European Conference on Computer Vision (ECCV), Part I*, pp. 766-779, Marseille, France, Oct. 2008.
- [Celikcan13] U. Celikcan, G. Cimen, E. B. Kevinc, and T. Capin. Attention-aware disparity control in interactive environments. *Visual Computer*, vol. 29, issues 6–8, pp. 685–694, 2013.
- [Chapiro14] A. Chapiro, S. Heinzle, T. O. Aydin, S. Poulakos, M. Zwicker, A. Smolic, and M. Gross. Optimizing Stereo-to-Multiview Conversion for Autostereoscopic Displays. *Computer Graphics Forum (Proc. Eurographics)*, vol. 33, no. 2, Strasbourg, France, April 2014.
- [Cheng08] C.-M. Cheng, S.-J. Lin, S.-H. Lai, and J.-C. Yang. Improved novel view synthesis from depth image with large baseline. In *International Conference on Pattern Recognition (ICPR)*, Tampa, FL, USA, Dec. 2008.
- [Cox96] I. J. Cox, S. L. Hingorani, S. B. Rao, and B. M. Maggs. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding (CVIU)*, vol. 63, issue 3, pp. 542-567, 1996.
- [Cutting97] J. E. Cutting. How the Eye Measures Reality and Virtual Reality. *Behavior Research Methods, Instruments, and Computers*, vol. 29, issue 1, pp. 27-36, Feb. 1997.
- [Dashwood10] T. Dashwood. A Beginner’s Guide to Shooting Stereoscopic 3D. Online: <http://www.dashwood3d.com/blog/beginners-guide-to-shooting-stereoscopic-3d/>, May 2010.
- [Divorra10] O. Divorra, J. Civit, F. Zuo, H. Belt, I. Feldmann, O. Schreer, E. Yellin, W. Ijsselsteijn, R. van Eijk, D. Espinola, P. Hagendorf, W. Waizenegger, and R. Braspenning. Towards 3D-Aware Telepresence: Working on Technologies Behind the Scene. In *ACM Conf. on Computer Supported Cooperative Work (CSCW)*, New Frontiers in Telepresence, Savannah, Georgia, USA, Feb. 2010.
- [Dodgson05] N. A. Dodgson. Autostereoscopic 3D displays. *IEEE Computer*, vol. 38, issue 8, pp. 31-36, Aug. 2005.
- [Dumbreck98] A. Dumbreck, T. Alpert, B. Choquet, C. W. Smith, J. Fournier, and P. M. Scheiwiller. Stereo Camera Human Factors Specification. *DISTIMA Technical Report D15*, CEC-RACE-DISTIMA-R2045, 1998.
- [Ebrahimi09] M. Ebrahimi and W. Mayol-Cuevas. SUSurE: Speeded up surround extrema feature detector and descriptor for realtime applications. In *CVPR Workshops, Workshop on feature detectors and descriptors: the state of the art and beyond*, 2009.
- [Faugeras93a] O. Faugeras, B. Hotz, H. Matthieu, T. Vieville, Z. Zhang, P. Fua, E. Theron, L. Moll, G. Berry, J. Vuillemin, P. Bertin, and C. Proy. Real Time Correlation-Based Stereo: Algorithm, Implementations and Applications. *INRIA Technical Report 2013*, 1993.
- [Faugeras93b] O. Faugeras. *Three-Dimensional Computer Vision (Artificial Intelligence)*. The MIT Press, Nov. 1993.
- [Fehn02] C. Fehn, P. Kauff, M. O. de Beeck, F. Ernst, W. Ijsselsteijn, M. Pollefeys, L.V. Gool, E. Ofek, and I. Sexton. An evolutionary and optimised approach on 3D-TV. In *International Broadcasting Convention (IBC)*, pp. 357-365, Amsterdam, NL, 2002.
- [Fehn03a] C. Fehn. A 3D-TV approach using depth-image-based rendering (DIBR). In *IASTED International Conference on Visualization, Imaging and Image Processing (VIIP)*, Benalmadena, Spain, Sept. 2003.
- [Fehn03b] C. Fehn. A 3D-TV system based on video plus depth information. In *37th Asilomar Conference on Signals, Systems and Computers (ACSSC)*, vol. 2, pp. 1529-1533, Pacific Grove, CA, USA, Nov. 2003.
- [Fehn04] C. Fehn. Depth-Image Based Rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV. *Proc. SPIE, Stereoscopic Display and Virtual Reality Systems XI*, vol. 5291, pp. 93-104, May 2004.
- [Feldmann09a] I. Feldmann, O. Schreer, P. Kauff, R. Schäfer, Z. Fei, H.J.W. Belt, and Ò. Divorra Escoda. Immersive Multi-User 3D Video Communication. *International Broadcasting Convention (IBC)*, Amsterdam, NL, Sept. 2009.

- [Feldmann09b] I. Feldmann, N. Atzpadin, O. Schreer, J.-C. Pujol-Acolado, J.-L. Landabaso, and O. Divorra Escoda. Multi-View Depth Estimation Based on Visual-Hull Enhanced Hybrid Recursive Matching for 3D Video Conference Systems. In *International Conference on Image Processing (ICIP)*, pp. 745-748, Cairo, Egypt, Nov. 2009.
- [Fischler80] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting applications to image analysis and automated cartography. *Proc. Image Understanding Workshop*, pp. 71-88, April 1980.
- [Fobker11] D. Fobker. Objektive in der 3D-Broadcast-Live-Produktion. *Fachzeitschrift für Fernsehen, Film und Elektronische Medien (FKT)*, 5/2011, pp. 222-225, 2011.
- [Foessel09] S. Foessel. 3D-Wiedergabe im Kino. *Fachzeitschrift für Fernsehen, Film und Elektronische Medien (FKT)*, 4/2009, pp. 170-173, 2009.
- [Forstmann04] S. Forstmann, Y. Kanou, Jun Ohya, S. Thuering, A. Schmitt. Real-Time Stereo by using Dynamic Programming. In *Computer Vision and Pattern Recognition Workshop, 2004. (CVPRW 2004)*, pp. 29, June/July 2004.
- [Fraser06] C. S. Fraser and S. Al-Ajlouni. Zoom-dependent camera calibration in digital close-range photogrammetry. *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 9, pp. 1017-1026, Sept. 2006.
- [Fusiello00] A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, vol. 12, issue 1, pp. 16-22, July 2000.
- [Fusiello08] A. Fusiello and L. Irsara. Quasi-euclidean uncalibrated epipolar rectification. In *International Conference on Pattern Recognition (ICPR)*, Tampa, FL, USA, Dec. 2008.
- [Georgiev13] M. Georgiev, A. Gotchev, and M. Hannuksela. A fast and accurate re-calibration technique for misaligned stereo cameras. In *International Conference on Image Processing (ICIP)*, pp. 24-28, Melbourne, Australia, Sept. 2013.
- [Giardina12] C. Giardina. NAB2012: Producer Jon Landau Reveals How ‘Titanic’ Was Converted to 3D. *The Hollywood Reporter*, 17th of April 2012.
- [Grau09] O. Grau and J. Kluger. Sport-Events in 3D. *Fachzeitschrift für Fernsehen, Film und Elektronische Medien (FKT)*, 4/2011, pp. 152-157, 2011.
- [Grau11] O. Grau, T. Borel, P. Kauff, A. Smolic, and R. Tanger. 3D-TV R&D Activities in Europe. *IEEE Trans. on Broadcasting*, vol. 57, issue 2, part 2, pp. 408-420, June 2011.
- [Greisen11] P. Greisen, S. Heinzle, M. Gross, and A. P. Burg. An FPGA-based processing pipeline for high-definition stereo video. *EURASIP Journal on Image and Video Processing*, 2011:18, Nov. 2011.
- [Harris88] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference (AVC)*, pp. 147-151, Manchester, UK, Aug./Sept. 1988.
- [Hartley99] R. I. Hartley. Theory and practice of projective rectification. *International Journal of Computer Vision (IJCV)*, vol. 35, issue 2, pp. 115-127, Nov. 1999.
- [Hartley04] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.
- [He10] W. He, W. Guozhong, L. Liliang, Z. Yang, A. Ping, and Z. Zhaoyang. Fast automatic elimination of vertical parallax of multiview images. *10th Intern. Conf. on Signal Processing (ICSP)*, pp. 1004-1007, Beijing, China, Oct. 2010.
- [Heinrichs06] M. Heinrichs and V. Rodehorst. Trinocular Rectification For Various Camera Setups. In *Symposium of International Society for Photogrammetry and Remote Sensing (ISPRS), Comm. III, PCV’06*, pp. 43-48, Bonn, Germany, Sept. 2006.
- [Heinzle11] S. Heinzle, P. Greisen, D. Gallup, C. Chen, D. Saner, A. Smolic, A. Burg, W. Matusik, and M. Gross. Computational Stereo Camera System with Programmable Control Loop. *ACM Trans. on Graphics (TOG) – Proc. of ACM SIGGRAPH 2011*, vol. 30 issue 4, no. 94, pp. 1-10, Vancouver, Canada, July 2011.
- [Hirschmüller08] H. Hirschmüller. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, issue 2, pp. 328-341, Feb. 2008.
- [Ho10] Y.-S. Ho. Recent Activities for 3DTV Research. In *Proc. of International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pp. 2-5, Pattaya, Thailand, July 2010.
- [Holliman04] N. Holliman. Mapping Perceived Depth to Regions of Interest in Stereoscopic Images. *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291, pp. 117-128, May 2004.
- [Holliman11] N. S. Holliman, N. A. Dodgson, G. E. Favalora, and L. Pockett. Three-dimensional displays: A review and applications analysis. *IEEE Trans. on Broadcasting*, vol. 57, issue 2, part 2, pp. 362-371, June 2011.
- [Iddan01] G. J. Iddan and G. Yahav. Three-dimensional imaging in the studio and elsewhere. *Proc. SPIE, Three-Dimensional Image Capture and Applications IV*, vol. 4298, pp. 48-55, April 2001.

- [IJsselsteijn00] W. A. IJsselsteijn, H. de Ridder, and J. Vliegen. Effects of Stereoscopic Filming Parameters and Display Duration on the Subjective Assessment of Eye Strain. *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems VII*, vol. 3957, pp. 12-22, May 2000.
- [IJsselsteijn02] W. A. IJsselsteijn, P. J. H. Seuntjens, and L. M. J. Meesters. State-of-the-Art in Human Factors and Quality Issues of Stereoscopic Broadcast Television. *ATTEST Technical Report D1*, IST-2001-34396, Aug. 2002.
- [Intel07] Intel Corporation. Intel SSE4 Programming Reference. Reference Number: D91561-001, April 2007.
- [Isgro99] F. Isgro and E. Trucco. Projective rectification without epipolar geometry. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1094-1099, 1999.
- [Iwasawa11] S. Iwasawa, M. Kawakita, S. Yano, M. Sakai, Y. Haino, M. Sato, and N. Inoue. A 200-Inch 3D-Glasses-Free High-Definition Projection Display. *SMPTE Conf. Proc.*, Oct. 2011.
- [Jones01] G. Jones, D. Lee, N. Holliman, and D. Ezra. Controlling Perceived Depth in Stereoscopic Images. *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems VIII*, vol. 4297, pp. 42-53, June 2001.
- [Kadir04] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. *European Conference on Computer Vision (ECCV)*, pp. 404-416, 2004.
- [Kang08] Y. Kang, C. Lee, and Y. Ho. An efficient rectification algorithm for multi-view images in parallel camera array. *3DTV-Conference (3DTV-CON)*, pp. 61-64, Istanbul, Turkey, May 2008.
- [Kangni06] F. Kangni, R. Laganier. Projective Rectification of Image Triplets from the Fundamental Matrix. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. II, pp. 14-19 Toulouse, France, May 2006.
- [Kauff07] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger. Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability. *Signal Processing: Image Communication*, vol. 22, issue 2, pp. 217-234, Feb. 2007.
- [Ke04] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pages 506-513, Washington, DC, USA, June/July 2004.
- [Kim13] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. on Graphics (TOG) – Proc. of ACM SIGGRAPH 2013*, vol. 32, issue 4, no. 73, pp. 1-12, Anaheim, USA, July 2013.
- [Knorr06] S. Knorr, E. Imre, B. Ozkalayci, A. A. Alatan, and T. Sikora. A Modular Scheme for 2D/3D Conversion of TV Broadcast. In *3rd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, pp. 703-710, Chapel Hill, USA, June 2006.
- [Knorr12] S. Knorr, K. Ide, M. Kunter, and T. Sikora. The Avoidance of Visual Discomfort and Basic Rules for Producing ‘Good 3D’ Pictures. *SMPTE Motion Imaging Journal*, vol. 121, issue 7, pp. 72-79, Oct. 2012.
- [Köppel10] M. Köppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, T. Wiegand. Temporally Consistent Handling of Disocclusions with Texture Synthesis for Depth-Image-based Rendering. In *International Conference on Image Processing (ICIP)*, pp. 1809-1812, Hong Kong, Sept. 2010.
- [Kopf07] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele. Joint Bilateral Upsampling. *ACM Trans. on Graphics (TOG) – Proc. of ACM SIGGRAPH 2007*, vol. 26, issue 3, no. 96, pp. 1-5, San Diego, USA, 2007.
- [Kurillo13] G. Kurillo, H. Baker, Z. Li, and R. Bajcsy. Geometric and Color Calibration of Multiview Panoramic Cameras for Life-Size 3D Immersive Video. In *International Conference on 3D Vision (3DV)*, pp. 374-381, Seattle, WA, June/July 2013.
- [Lambooi09] M. T. M. Lambooi, W.A. IJsselsteijn, M. Fortuin, I. Heyndericks. Visual Discomfort and Visual Fatigue of Stereoscopic Displays: A Review. *Journal of Imaging Science and Technology*, vol. 53, issue 3, pp. 030201– 030201-14, 2009.
- [Lang10] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. Nonlinear disparity mapping for stereoscopic 3D. *ACM Trans. on Graphics (TOG) – Proc. of ACM SIGGRAPH 2010*, vol. 29, issue 4, no. 75, pp. 1-10, Los Angeles, USA, July 2010.
- [Lewis95] J. P. Lewis. Fast Template Matching. *Vision Interface*, pp. 120-123, 1995.
- [Lindeberg98] T. Lindeberg. Feature Detection with Automatic Scale Selection. *Intern. J. Computer Vision*, vol. 30, no. 2, pp. 79-116, 1998.
- [Lipton82] L. Lipton, *Foundations of the Stereoscopic Cinema – A Study in Depth*, Van Nostrand Reinhold, New York, NY, USA, 1982.

- [Lipton97] L. Lipton, *StereoGraphics Developers' Handbook*, StereoGraphics Corporation, 1997.
- [Lipton01] L. Lipton. The Stereoscopic Cinema: From Film to Digital Projection. *SMPTE Journal*, pp. 586-593, Sept. 2001.
- [Loop99] C. Loop and Z. Zhang. Computing rectifying homographies for stereo vision. *Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 131, vol. 1, 1999.
- [Lowe99] D.G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1150-1157, Kerkira, Greece, 1999.
- [Lowe04] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, vol. 60, issue 2, pp. 91-110, Nov. 2004.
- [Mallon05] J. Mallon, P. F. Whelan. Projective rectification from the fundamental matrix. *Image and Vision Computing*, vol. 23, issue 7, pp. 643-650, July 2005.
- [Matas04] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, vol. 22, issue 10, pp. 761-767, 2004.
- [Matusik04] W. Matusik and H. Pfister. 3D TV: A scalable system for real-time acquisition, transmission and autostereoscopic display of dynamic scenes. *ACM Trans. on Graphics (TOG) – Proc. of ACM SIGGRAPH 2004*, vol. 23, issue 3, pp. 814-824, Los Angeles, USA, Aug. 2004.
- [Mendiburu08] B. Mendiburu. *3D Movie Making – Stereoscopic Digital Cinema from Script to Screen*. Elsevier, ISBN: 978-0-240-81137-6, 2008.
- [Mendiburu12] B. Mendiburu, Yves Pupulin, and Steve Schklair, *3D TV and 3D Cinema: Tools and Processes for Creative Stereoscapy*, Waltham, MA: Focal Press/Elsevier, 2012.
- [Mikolajczyk04] K. Mikolajczyk and C. Schmid. Scale and Affine invariant interest point detectors. *International Journal of Computer Vision (IJCV)*, vol. 60, issue 1, pp. 63-86, Oct. 2004.
- [Mikolajczyk05a] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27, issue 10, pp. 1615-1630, Oct. 2005.
- [Mikolajczyk05b] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision (IJCV)*, vol. 65, issue 1-2, pp. 43-72, Nov. 2005.
- [Imagineer] Imagineer Systems. Mocha Pro 4.1. <http://www.imagineersystems.com/>
- [MüllerK08] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand. View Synthesis for Advanced 3D Video Systems. *EURASIP J. on Image and Video Proc.*, 2008.
- [MüllerK11] K. Müller, P. Merkle, and T. Wiegand. 3D Video Representation Using Depth Maps. *Proceedings of the IEEE (PIEEE)*, 2011.
- [Muscade] Muscade (MULTImedia SCALable 3D for Europe). *Seventh Framework Theme ICT-2009.1.5 Networked Media and 3D Internet*. Grant Agreement no. 247010. <http://www.muscade.eu/>
- [Nagano13] K. Nagano, A. Jones, J. Liu, J. Busch, X. Yu, M. Bolas, and P. Debevec. An autostereoscopic projector array optimized for 3D facial display. *ACM SIGGRAPH 2013 Emerging Technologies*, no. 3, Anaheim, CA, USA, July 2013.
- [Ndjiki-Nya10] P. Ndjiki-Nya, M. Köppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand. Depth-Image Based Rendering with Advanced Texture Synthesis. In *Proc. IEEE Intern. Conference on Multimedia & Expo (ICME)*, pp. 424-429, Suntec City, Singapore, 2010.
- [Neubeck06] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *International Conference on Pattern Recognition (ICPR)*, pp. 850-855, Hong Kong, Aug. 2006.
- [Nozick11] V. Nozick. Multiple view image rectification. In *1st Intern. Symposium on Access Spaces (ISAS)*, pp. 277-282, June 2011.
- [Okutomi93] M. Okutomi and T. Kanade. A Multiple-Baseline Stereo. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 15, issue 4, pp. 353-363, April 1993.
- [Oskam11] T. Oskam, A. Hornung, H. Bowles, K. Mitchell, M. Gross. OSCAM - optimized stereoscopic camera control for interactive 3D. *ACM Trans. on Graphics (TOG) – Proc. of ACM SIGGRAPH Asia 2011*, vol. 30, issue 6, no. 189, pp. 1-8, Hong Kong, Dec. 2011.
- [Papadimitriou96] D. Papadimitriou and T. Dennis. Epipolar line estimation and rectification for stereo image pairs. *IEEE Trans. on Image Processing*, vol. 5, issue 4, pp. 672-676, April 1996.
- [Pastoor95] S. Pastoor. Human Factors of 3D Imaging: Results of Recent Research at Heinrich-Hertz-Institut Berlin. In *2nd International Display Workshop (IDW)*, Hamamatsu, Japan, Oct. 1995.

- [Redert02] A. Redert, M. Op de Beeck, C. Fehn, W. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, I. Sexton, and P. Surman. ATTEST – Advanced Three-Dimensional Television Systems Technologies. In *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, pp. 313-319, Padova, Italy, June 2002.
- [Riemens09] A. K. Riemens, O. P. Gangwal, B. Barenbrug, and R.-P. M. Berretty. Multi-step joint bilateral depth upsampling. *Proc. SPIE, Visual Communications and Image Processing*, vol. 7257, pp. 1-12, 2009.
- [Routier12] P. Routhier and T. Borel. certifie3D – 3D Quality Analysis Service. Certifi3D datasheet, Technicolor, 2012.
- [Schaffner13] M. Schaffner, P. Hager, L. Cavigelli, P. Greisen, F.K. Gurkaynak, and H. Kaeslin. A real-time 720p feature extraction core based on Semantic Kernels Binarized. In *IFIP/IEEE 21st International Conf. on Very Large Scale Integration (VLSI-SoC)*, pp. 27-32, Istanbul, Turkey, Oct. 2013.
- [Scharstein02] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, vol. 47, issue 1-3, pp. 7-42, April-June 2002. Microsoft Research Technical Report MSR-TR-2001-81, Nov. 2001.
- [Schmidt11] W. Schmidt. Wege in der 3D-Liveproduktion. *Fachzeitschrift für Fernsehen, Film und Elektronische Medien (FKT)*, 8-9/2011, p. 462, 2011.
- [Schreer08] O. Schreer, I. Feldmann, N. Atzpadin, P. Eisert, P. Kauff, and H. J. W. Belt. 3DPresence - A System Concept for Multi-User and Multi-Party Immersive 3D Videoconferencing. In *Conference on Visual Media Production (CVMP)*, pp. 1-8, London, UK, Nov. 2008.
- [Shashua95] A. Shashua. Algebraic functions for recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol.17, issue 8, pp. 779-789, Aug. 1995.
- [Shi94] J. Shi, C. Tomasi. Good features to track. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 593-600, Seattle, USA, 1994.
- [SilhouetteFX] SilhouetteFX, LLC. Silhouette fx v5. <http://www.silhouetefx.com/>
- [Smolic08] A. Smolic, K. Müller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand. Intermediate View Interpolation based on Multiview Video plus Depth for Advanced 3D Video Systems. In *International Conference on Image Processing (ICIP)*, pp. 2448-2451, Oct. 2008.
- [Smolic11a] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Müller, and M. Lang. Three-Dimensional Video Postproduction and Processing. *Proceedings of the IEEE (PIEEE)*, April 2011.
- [Smolic11b] A. Smolic. 3D video and free viewpoint video—From capture to display. *Pattern Recognition*, vol. 44, issue 9, pp. 1958-1968, ISSN 0031-3203, Sept. 2011.
- [Sony] Sony. MPE 200, Multi Image Processor. <http://pro.sony.com/>
- [Sony12] Sony. 3D at 2D economics. Sony Pictures Technologies, *White Paper*, Aug. 2012.
- [Stefanoski13] N. Stefanoski, O. Wang, M. Lang, P. Greisen, S. Heinzle, and A. Smolic. Automatic View Synthesis by Image-Domain-Warping. *IEEE Trans. on Image Processing*, vol. 22, no. 9, pp. 3329-3341, Sept. 2013.
- [Stereolabs] Stereolabs. PURE ON-SET, 3D Monitoring and Video Assist. <http://www.stereolabs.com>
- [Stommel10] M. Stommel. Binarising SIFT-Descriptors to Reduce the Curse of Dimensionality in Histogram-Based Object Recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 3, no. 1, March 2010.
- [Sun03] C. Sun. Uncalibrated three-view image rectification. *Image and Vision Computing*, vol. 21, issue 3, pp. 259-269. March 2003.
- [Tanger13] R. Tanger, M. Müller, P. Kauff, R. Schäfer. Depth/Disparity Creation for Trifocal Hybrid 3D System. *SMPTE Conference Proceedings*, Oct. 2013.
- [Tanimoto06] M. Tanimoto. Overview of free viewpoint television. *Signal Processing: Image Communication*, 21 (2006) 454–461, 2006.
- [Templin14] K. Templin, P. Didyk, K. Myskowski, M. M. Hefeeda, H.-P. Seidel, and W. Matusik. Modeling and optimizing eye vergence response to stereoscopic cuts. *ACM Trans. on Graphics (TOG) – Proc. of ACM SIGGRAPH 2014*, vol. 33, issue 4, no. 145, Vancouver, Canada, July 2014.
- [Tomasi98] C. Tomasi and R. Manduchi. Bilateral Filtering for Gray and Color Images. In *International Conference on Computer Vision (ICCV)*, pp. 839-846, Jan. 1998.
- [Tuytelaars04] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision (IJCV)*, vol. 59, issue 1, pp. 61-85, Aug. 2004.

- [Ventura11] J. Ventura and T. Höllerer. Fast and Scalable Keypoint Recognition and Image Retrieval using Binary Codes. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 697-702, Kona, HI, USA, Jan. 2011.
- [Viola01] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 511-518, Kauai, HI, USA, Dec. 2001.
- [Wagner10] R. Wagner. 3D-Live-Konzert-Übertragung aus dem Steintor Varieté in Halle/Saale. *Fachzeitschrift für Fernsehen, Film und Elektronische Medien (FKT)*, 11/2010, pp. 583-584, 2010.
- [Waizenegger11] W. Waizenegger, I. Feldmann, and O. Schreer. Real-time Patch Sweeping for High-Quality Depth Estimation in 3D Videoconferencing Applications. *Proc. SPIE, Real-Time Image and Video Processing*, vol. 7871, pp. 1-10, *Invited Paper*, Feb. 2011.
- [Wheatstone38] C. Wheatstone. Contributions to the Physiology of Vision.—Part the First. On some remarkable, and hitherto unobserved, Phenomena of Binocular Vision. *Philosophical Transactions of the Royal Society of London*, vol. 128, pp. 371-394. Received and Read June 21, 1838.
- [Wilburn05] B. Wilburn, N. Joshi, V. Vaish, E. Talvala, E. Antunez, A. Barth, A. Adamas, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Trans. on Graphics (TOG) – Proc. of ACM SIGGRAPH 2005*, vol. 24, issue 3, pp. 765-776, Los Angeles, USA, July 2005.
- [Woods93] A. Woods, T. Docherty, and R. Koch. Image Distortions in Stereoscopic Video Systems. *Proc. SPIE, Stereoscopic Displays and Applications IV*, vol. 1915, pp. 36-48, Sept. 1993.
- [Woods14] A. Woods. The 3D Movie List. <http://www.3dmovielist.com/>, Last updated 2014.
- [Wopking95] M. Wopking. Viewing comfort with stereoscopic pictures: an experimental study on the subjective effects of disparity magnitude and depth of focus. *Journal of the Society for Information Display (SID)*, vol. 3, pp. 101-103, 1995.
- [Wu05] H.-H. Wu and Y.-H. Yu. Projective rectification with reduced geometric distortion for stereo vision and stereoscopic video. *Journal of Intelligent and Robotic Systems*, vol. 42, issue 1, pp. 71-94, Jan. 2005.
- [Wu13] B. Wu, H. Hu, Q. Zhu, and Y. Zhang. A Flexible Method for Zoom Lens Calibration and Modeling Using a Planar Checkerboard. *Photogrammetric Engineering & Remote Sensing*, vol. 79, no. 6, pp. 555-571, June 2013.
- [Xin04] Du Xin and Li Hongdong. A Simple Rectification Method for Linear Multi-baseline Stereovision System. *Journal of Zhejiang University (Science)*, vol. 5, issue 5, pp. 567-571, June 2004.
- [Yang10] Z. Yang, A. Ping, W. He, and Z. Zhaoyang. A rectification algorithm for un-calibrated multi-view images based on SIFT features. In *International Conference on Audio Language and Image Processing (ICALIP)*, pp. 143-147, Nov. 2010.
- [Zhang95] Z. Zhang, R. Deriche, O. D. Faugeras, and Q. T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, vol. 78, issues 1-2, pp. 87-119, Oct. 1995.
- [Zhang03] H. Zhang, J. Cech, R. Sara, F. Wu and Z. Hu. A Linear Trinocular Rectification Method for Accurate Stereoscopic Matching. In *British Machine Vision Conference (BMVC)*, pp. 29.1-10, Norwich, UK, Sept. 2003.
- [Zhang07] G. Zhang, W. Hua, X. Qin, T.-T. Wong, H. Bao. Stereoscopic Video Synthesis from a Monocular Video. *IEEE Trans. on Visualization and Computer Graphics*, vol. 13, no. 4, pp. 686-696, Jul./Aug. 2007.
- [ZhangL11] L. Zhang, C. Vazquez and S. Knorr. 3D-TV content creation: Automatic 2d-to-3d video conversion. *IEEE Trans. on Broadcasting*, vol. 57, issue 2, part 2, pp. 372-383, June 2011.
- [ZhangQ11] Q. Zhang, P. An, Y. Zhang, L. Shen, and Z. Zhang. Improved multi-view depth estimation for view synthesis in 3D video coding. In *3DTV-Conference (3DTV-CON)*, Antalya, Turkey, May 2011.
- [Zitnick04] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. on Graphics (TOG) - Proc. of ACM SIGGRAPH 2004*, vol. 23, issue 3, pp. 600-608, Los Angeles, USA, Aug. 2004.