Sebastian Knorr

Synthese stereoskopischer Sequenzen aus 2-dimensionalen Videoaufnahmen

Synthese stereoskopischer Sequenzen aus 2-dimensionalen Videoaufnahmen

Von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
– Dr.-Ing. –
genehmigte Dissertation

vorgelegt von
Dipl.-Ing. Sebastian Knorr

Promotionsausschuss

Vorsitzender : Prof. Dr. F. Wichmann
1. Gutachter: Prof. Dr.-Ing. T. Sikora
2. Gutachter: Prof. Dr.-Ing. J. Ostermann

Tag der wissenschaftlichen Aussprache: 27.10.2008

Berlin 2008
D 83

Danksagung

Diese Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Fachgebiet Nachrichtenübertragung der TU Berlin. An dieser Stelle möchte ich mich bei allen bedanken, die mich auf diesem langen und nicht immer sehr einfachen Weg unterstützt haben.

Mein erster Dank geht an Herrn Prof. Dr.-Ing. Thomas Sikora, der mich als wissenschaftlicher Mitarbeiter auch ohne jegliche Vorkenntnisse in der Bildsignalverarbeitung in seiner Abteilung aufnahm und mir ein doch sehr interessantes Forschungsthema vorschlug. Für die hervorragende Zusammenarbeit möchte ich mich des weiteren bei meinen Kollegen bedanken, insbesondere bei meinem jetzigen Geschäftspartner Dr. Matthias Kunter, durch dessen eigene Arbeit und Anregungen erst viele neue Ideen entstanden sind, sowie bei Andreas Krutz, Engin Kurutepe und Carsten Clemens für das Korrekturlesen. Ferner möchte ich mich bei meinem Diplomarbeiter Thomas Walter bedanken, der mich bei Teilen meiner Arbeit sehr unterstützt hat.

Sehr geholfen hat mir auch die Beteiligung an dem EU-Projekt „3DTV“, durch das ich einige Kontakte zu anderen Wissenschaftlern in meinem Forschungsbereich aufbauen konnte. Dabei möchte ich besonders Evren Imre, Engin Tola und Prof. Dr. Aydin A. Alatan für eine tolle Zusammenarbeit danken. Ein sehr großer Dank gilt auch den Kollegen Ingo Feldmann, Dr. Christoph Fehn, Peter Kauff, Dr. Peter Eisert, Dr. Aljoscha Smolic und Dr. Karsten Müller vom Heinrich-Hertz-Institut (HHI) für die vielen Diskussionen und Anregungen.

Des weiteren möchte ich meinen Kollegen Dr. Matthias Kunter, Carsten Clemens, Marcus Schwab, Dr. Zouhair Belkoura, Dr. Jan-Mark Batke, Tilman Liebchen, Jan Weil, Engin Kurutepe, Andreas Krutz, Kai Clüver, Rüdiger Knörig und Georges Haboub für die lustigen Zeiten am Institut danken. Ein ganz lieber Dank richtet sich auch an Birgit Boldin, die in dem Verwaltungschaos der TU Berlin immer den Durchblick hatte. Sie ist das Herz und die Seele des Fachgebiets.

Für das Korrekturlesen der gesamten Arbeit möchte ich mich bei meiner Schwiegermutter Petra Hempke bedanken.

Mein ganz besonderer und herzlicher Dank gilt meiner Ehefrau Nadine Hempke, die ich sehr liebe. Sie hat mich über all die Jahre immer unterstützt und mir liebevoll zur Seite gestanden.

Eidesstattliche Versicherung

Ich versichere an Eides Statt, dass ich die von mir vorgelegte Dissertation selbstständig angefertigt und alle benutzten Quellen und Hilfsmittel vollständig angegeben habe.

Eine Anmeldung der Promotionsabsicht habe ich an keiner anderen Fakultät oder Hochschule beantragt.

Kurzfassung

Die rasante Entwicklung der 3D-Technologie führt derzeit zu einer verstärkten Nachfrage nach 3D-Spielen, 3D-Filmen, 3D-Präsentationen und interaktiven 3D-Inhalten. Von der Motiviation getrieben, alte Filme als Neuauflage in 3D zu erleben, beschäftigt sich die vorliegende Dissertation mit der automatischen Konvertierung von Videosequenzen von 2D in stereoskopisch 3D. Zur Umsetzung dieses Ziels wird ein neues Verfahren zur realistischen Synthese stereoskopischer Ansichten (RSVS) aus monokularen Videosequenzen, welches auf Ansätzen aus den Forschungsbereichen der Computer Vision (CV) und Computergrafik (CG) basiert, vorgestellt. Die Ergebnisse von RSVS werden anhand zahlreicher Sequenzen mit Verfahren des aktuellen Forschungsstandes verglichen und zeigen die hervorragenden Konvertierungseigenschaften dieses Ansatzes.

Ferner werden weitere Ansätze beschrieben, die den Einsatz von RSVS erweitern. Zum einen wird gezeigt, dass RSVS nicht nur stereoskopische Sequenzen aus monokularen Videosequenzen generiert, sondern auch in der Lage ist, Multiview-Video für Multi-user 3D-Displays zu erzeugen. Zum anderen lässt sich RSVS mit Super-Resolutionsansätzen erweitern und somit sowohl die Qualität des Originalvideos als auch der synthetisierten stereoskopischen Ansichten erhöhen.

Inhaltsverzeichnis

Eidesstattliche Versicherung **vii**

1. Einleitung **1**
1.1. Motivation . . . 1
1.2. Einführung in die Stereoskopie . . . 3
1.2.1. Physiologie des 3-D Sehens . . . 3
1.2.2. Geschichte der Stereoskopie . . . 6
1.3. Stand der Forschung . . . 8
1.3.1. Tiefenschätzung aus Bildern und Bildfolgen . . . 8
1.3.2. Synthese virtueller Ansichten . . . 9
1.3.3. Darstellungsmethoden von stereoskopischen Ansichten . . . 10
1.4. Überblick über diese Arbeit . . . 13
1.4.1. Eigene Ansätze . . . 13
1.4.2. Wesentliche Beiträge und neue Ansätze . . . 15
1.4.3. Aufbau der Arbeit . . . 15

2. Theoretische Vorbetrachtung **17**
2.1. Projektive Geometrie . . . 17
2.1.1. Homogene Darstellung von Punkten und Linien in $\mathbb{P}^2$ . . . 19
2.2. Kameramodell . . . 19
2.2.1. Lochkamera und Zentralprojektion . . . 20
2.2.2. Verallgemeinerte Zentralprojektion . . . 21
2.3. Epipolargeometrie . . . 21
2.3.1. Fundamentalmatrix . . . 22
2.3.2. Essentialmatrix . . . 24
2.3.3. Eigenschaften der Fundamental- und Essentialmatrix . . . 24
2.4. Homographie zwischen zwei Ansichten . . . 25
2.5. Kapitelzusammenfassung . . . 26

3. Robuste Schätzverfahren **27**
3.1. Lineare Schätzverfahren . . . 28
3.1.1. 8-Punkt-Algorithmus . . . 28

3.1.2. Normalisierter 8-Punkt-Algorithmus . . . 30
3.2. Nichtlineare Schätzverfahren . . . 31
3.2.1. Gradientenverfahren . . . 33
3.2.2. Gauß-Newton-Verfahren . . . 33
3.2.3. Levenberg-Marquardt Verfahren . . . 34
3.3. Robuste Schätzverfahren . . . 35
3.3.1. Monte-Carlo-Methoden . . . 36
3.3.2. Fehlermaße . . . 40
3.4. Vergleich der robusten Schätzverfahren . . . 41
3.4.1. Modellierung der Epipolargeometrie . . . 41
3.4.2. Fehlermaße zum Vergleich der Ergebnisse . . . 42
3.4.3. Ergebnisse und Auswertung . . . 42
3.5. Kapitelzusammenfassung und Diskussion . . . 50

4. 3D-Rekonstruktion aus zwei Ansichten **53**
4.1. Problemstellung und Überblick über die Methode . . . 53
4.2. Korrespondenzanalyse . . . 55
4.2.1. Merkmalspunktdetektor . . . 56
4.2.2. Suche nach eindeutigen Zuordnungen . . . 59
4.3. Schätzung der Epipolargeometrie . . . 65
4.3.1. Initialisierung durch Eliminierung von Ausreißern . . . 66
4.3.2. Optimierung der Schätzergebnisse . . . 69
4.3.3. Rektifizierung . . . 72
4.4. Selbstkalibrierung . . . 75
4.5. Bestimmung der Rotation und Translation . . . 77
4.6. Triangulation . . . 78
4.6.1. Mittelpunkt-Methode . . . 78
4.6.2. Optimale Triangulation . . . 79
4.7. Ergebnisse . . . 80
4.8. Kapitelzusammenfassung . . . 81

5. 3D-Rekonstruktion aus mehreren Ansichten **85**
5.1. Problemstellung und Überblick über die Methode . . . 85
5.2. Verfolgung von Merkmalspunkten . . . 87
5.2.1. Vorbetrachtung . . . 87
5.2.2. Affines Bewegungsmodell . . . 89
5.2.3. Iteratives Lucas-Kanade-Verfahren . . . 92
5.2.4. Pyramidenförmige Implementierung des iterativen Lucas-Kanade-Verfahrens . . . 94

5.2.5. Qualität der verfolgten Merkmalspunkte 99
5.2.6. Ergebnisse 103
5.3. Detektieren von Schlüsselbildern 108
5.4. Initiale Rekonstruktion 112
5.5. Hinzufügen weiterer Ansichten 113
5.5.1. Bestimmung der Kameraprojektionsmatrizen über 2D-3D-Punktkorrespondenzen 113
5.5.2. Bestimmung weiterer 3D-Raumpunkte mittels Triangulation 117
5.5.3. Optimierung der Kameraparameter und der 3D-Koordinaten der Raumpunkte 117
5.5.4. Implementierungsvarianten 120
5.5.5. Rekonstruktionsergebnisse 122
5.6. Weitere Rekonstruktionsmethoden 125
5.7. Ergebnisse 127
5.8. Kapitelzusammenfassung 128

6. Generierung stereoskopischer Ansichten 131
6.1. Stereo aus Video und Tiefe 131
6.1.1. Problemstellung und Überblick über die Methode 131
6.1.2. Erzeugung dichter Tiefenkarten 133
6.1.3. Stereobildsynthese mittels DIBR 137
6.1.4. Aktuelle 2D/3D-Konvertierungsverfahren basierend auf DIBR . . . 144
6.2. Realistisches Stereo durch 2-D Ebenentransformation 146
6.2.1. Problemstellung und Überblick über die Methode 147
6.2.2. Definition virtueller Kameras 149
6.2.3. Bestimmung der 2D-Homographien 151
6.2.4. Stereobildsynthese 152
6.2.5. Super-Resolution Stereobildsynthese 155
6.2.6. Ergebnisse 159
6.3. Kapitelzusammenfassung und Diskussion 163

7. Evaluierung 167
7.1. Überblick über die Evaluierungsmethoden 167
7.2. Subjektiver Qualitätstest I (SS-Methode) 168
7.2.1. Single-Stimulus Methode 169
7.2.2. Auswertung 169
7.3. Subjektiver Qualitätstest II (DSCQS-Methode) 172
7.3.1. Double-Stimulus-Continous-Quality-Scale Methode 174
7.3.2. Auswertung 174

7.4. Kapitelzusammenfassung und Diskussion 177

8. Einführung in den Umgang mit dynamischen Szenen 179
8.1. Problemstellung und Überblick 179
8.2. Ansatz zur 3D-Rekonstruktion dynamischer Szenen 181
8.2.1. Tracking und Detektion von Schlüsselbildern 181
8.2.2. Geometrische Segmentierung 183
8.3. Rekonstruktionsergebnisse 187
8.4. Kapitelzusammenfassung, Diskussion und Ausblick 188

9. Zusammenfassung und Ausblick 191
9.1. Zusammenfassung 191
9.2. Wesentliche Beiträge dieser Arbeit 193
9.3. Ausblick 194

A. Anhang 197
A.1. Robuste Schätzung der Fundamentalmatrix 197
A.2. Levenberg-Marquardt-Algorithmus 205
A.3. Verwendete Bilder und Sequenzen 206
A.3.1. Verwendete Bilder in Kapitel 3 und 4 206
A.3.2. Verwendete Videosequenzen in Kapitel 5, 6 und 7 208
A.3.3. Verwendete Sequenzen in Kapitel 8 212
A.4. Informationsbögen für die subjektiven Qualitätstests 213

. Literaturverzeichnis 215

Abbildungsverzeichnis

1.1. 2D/3D-Konvertierung von Filmen 2
1.2. Gesichtsfeld, Sehbahn und visueller Kortex 3
1.3. Binokulares Sehen 4
1.4. Monokulare Tiefenwahrnehmung I 5
1.5. Monokulare Tiefenwahrnehmung II 5
1.6. Monokulare Tiefenwahrnehmung III 6
1.7. Wheatstone's Spiegelstereoskop 7
1.8. Arten der Parallaxe 11
1.9. Widerspruch zwischen monokularer und stereoskopischer Tiefenwahrnehmung 12
1.10. Autostereoskopische Darstellungsmethoden 13
1.11. Überblick über das 2D/3D-Konvertierungssystem 14

2.1. Projektive Transformation in eine Bildebene 18
2.2. Kameramodelle 20
2.3. Veralgemeinerte Zemtralprojektion 22
2.4. Epipolargeometrie 23

3.1. Bestimmung einer Ausgleichsgraden 28
3.2. Singularitäts-Bedingung der Fundamentalmatrix 30
3.3. Vergleich robuster Schätzer bei verrauschten Datensätzen 43
3.4. Vergleich robuster Schätzer bei der Eliminierung von Ausreißern ($\sigma = 0,5$ Pixel) 46
3.5. Verlauf des mittleren relativen Fehlers zwischen dem korrekten Epipol und den geschätzten Epipolen 47
3.6. Vergleich robuster Schätzer bei der Eliminierung von Ausreißern ($\sigma = 2,0$ Pixel) 49

4.1. Abhängigkeit des zu erwartenden Fehlers bei der Triangulation in Abhängigkeit vom Abstand der Kamerazentren 54
4.2. Degeneration der 3D-Struktur 55
4.3. Eigenschaften von Merkmalsobjekten 56

4.4. Vergleich verschiedener Autokorrelationsverläufe 58
4.5. Verteilung von Merkmalspunkten 58
4.6. Zuordnung durch Korrelation 60
4.7. Zuordnung mithilfe benachbarter Merkmalspunkte 62
4.8. Bestimmung eindeutiger Zuordnungen 67
4.9. Modifikation des RANSAC-Algorithmus 68
4.10. Eliminierung von Ausreißern (Bildpaar *Halle*) 70
4.11. Eliminierung von Ausreißern (Bildpaar *Turm*) 71
4.12. Planaren Rektifizierung 73
4.13. Polare Rektifizierung 74
4.14. Prinzip der polaren Rektifizierung 74
4.15. Rektifizierungsergebnisse 76
4.16. Mittelpunkt-Methode 79
4.17. Optimale Triangulation 80
4.18. Darstellung der Reprojektionsfehler 81
4.19. Rekonstruktionsergebnis für das Bildpaar *Turm* 82
4.20. Rekonstruktionsergebnis für das Bildpaar *Rathaus* 83
4.21. Rekonstruktionsergebnis für das Bildpaar *Totem* 84

5.1. Rückprojektion in eine rotierende Kamera 86
5.2. Pyramidische Repräsentation 94
5.3. Beispiel für einen Merkmalspunkt an einer Tiefendiskontinuität 99
5.4. Qualitätsbestimmung eines Merkmalspunktes 102
5.5. Anzahl der Merkmalspunkte über die Sequenz bei verschiedenen Fehlerfunktionen und Schwellwerten 104
5.6. Zunehmende Verdeckung eines Merkmalspunktes und den daraus resultierenden Fehlerwerten 106
5.7. Detektion von Schlüsselbildern 111
5.8. GRIC-Verläufe 112
5.9. Hinzufügen neuer Ansichten über bekannte 2D-3D-Korrespondenzen . . . 114
5.10. Mittlerer Reprojektionsfehler in den Schlüsselbildern 118
5.11. Bestimmung der Kameraparameter und 3D-Struktur 121
5.12. Vergleich des mittleren Reprojektionsfehlers für lineare Berechnung der Projektionsmatrizen 123
5.13. Vergleich des mittleren Reprojektionsfehlers für nichtlineare Berechnung der Projektionsmatrizen 124
5.14. Vergleich des mittleren Reprojektionsfehlers mit globalem Bundle Adjustment 124
5.15. Vergleich verschiedener Gewichtungsmaße zur Bestimmung neuer Bildpaare 127
5.16. Rekonstruktionsergebnis der Sequenz *Kirche* 128

5.17. Rekonstruktionsergebnis der Sequenz *Dom* 129

6.1. Darstellung von Video und Tiefe 132
6.2. Tiefenquatisierung 136
6.3. Generierung virtueller Ansichten 138
6.4. Shift-Sensor-Ansatz 140
6.5. Vorverarbeitung der Tiefenkarten 143
6.6. Stereobildsynthese mit Hintergrundextrapolation bzw. mit Gaußfilterung der Tiefenkarten 144
6.7. Ergebnisse der 2D/3D-Konvertierung nach [129] 145
6.8. Ergebnisse der 2D/3D-Konvertierung nach [93] 145
6.9. Ergebnisse der 2D/3D-Konvertierung nach [63] 146
6.10. Ergebnisse der 2D/3D-Konvertierung nach [152] 147
6.11. Überblick über das 2D/3D-Konvertierungssystem 148
6.12. Beispiel für RSVS 149
6.13. Filmauschnitt aus *Zurück in die Zukunft* 150
6.14. Bestimmung der 2D-Homographie und Stereobildsynthese 152
6.15. Stereosynthese mit umliegenden Originalansichten 154
6.16. Auffüllen des Randbereichs 155
6.17. Super-Resolution durch räumliche und zeitliche Redundanz 158
6.18. Anaglyph Stereobilder (Handkamera) 160
6.19. Anaglyph Stereobilder (BBC) 161
6.20. Multiview Ansichten (Handkamera) 161
6.21. Multiview Ansichten (BBC) 161
6.22. Standard RSVS vs. RSVS mit Super-Resolution (Statue) 162
6.23. Standard RSVS vs. RSVS mit Super-Resolution (Dom) 164

7.1. Beispielbilder für Qualitätstest I 169
7.2. Subjektiver Qualitätstest I 170
7.3. Beispielbilder für unterschiedliche Tiefenverteilung 172
7.4. Beispielbilder für Qualitätstest II 173
7.5. Zeitlicher Ablauf der DSCQS-Methode 174
7.6. Subjektiver Qualitätstest II 175

8.1. Vergleich der GRIC-Verläufe 182
8.2. Klassifikation von Trajektorienarten 183
8.3. Robuste IMO-Segmentierung 184
8.4. GRIC-Verläufe der synthetischen Sequenz *TUBroom-IMO* 185
8.5. Klassifizierung der Merkmalspunkte 186
8.6. Klassifizierung der Merkmalspunkte und 3D-Rekonstruktion 188

8.7. 3D-Rekonstruktion der *Sofa*-Sequenz 189

9.1. Ausblick: Semiautomatische 2D/3D-Konvertierung mittels RSVS 194

A.1. Bildpaar *Halle* 206
A.2. Bildpaar *Turm* 206
A.3. Bildpaar *Korridor* 207
A.4. Bildpaar *Rathaus* 207
A.5. Bildpaar *Totem* 207
A.6. Beispielbilder der Sequenz *TUBroom* 208
A.7. Beispielbilder der Sequenz *Palast* 208
A.8. Beispielbilder der Sequenz *Felsen* 208
A.9. Beispielbilder der Sequenz *Dom* 208
A.10. Beispielbilder der Sequenz *Kirche* 209
A.11. Beispielbilder der Sequenz *Medusa* 209
A.12. Beispielbilder der Sequenz *Fassade* 209
A.13. Beispielbilder der Sequenz *Statue* 209
A.14. Beispielbilder der Sequenz *Höhle* 210
A.15. Beispielbilder der Sequenz *Monument Valley* 210
A.16. Beispielbilder der Sequenz *Wald* 210
A.17. Beispielbilder der Sequenz *Riff* 210
A.18. Beispielbilder der Sequenz *Eisberg* 211
A.19. Beispielbilder der Sequenz *Dschungel* 211
A.20. Beispielbilder der Sequenz *Berge* 211
A.21. Beispielbilder der Sequenz *Tiefsee* 211
A.22. Beispielbilder der Sequenz *TUBroom-IMO* 212
A.23. Beispielbilder der Sequenz *Box* 212
A.24. Beispielbilder der Sequenz *Sofa* 212
A.25. Informationsbogen Qualitätstest I 213
A.26. Informationsbogen Qualitätstest II 214

1. Einleitung

3D has the opportunity to be the most important innovation in film since the use of color in motion pictures.

Jeffrey Katzenberg, CEO von DreamWorks Animation (Showest, 2005)

1.1. Motivation

Die rasante Entwicklung der 3D-Technologie führt derzeit zu einer verstärkten Nachfrage nach 3D-Spielen, 3D-Filmen, 3D-Präsentationen und interaktiven 3D-Inhalten. Viele Filmproduktionsstudios wie DreamWorks kündigten bereits an, all ihre zukünftigen Filme in 3D zu produzieren. Die großen Kinobetreiber wie UCI und Odeon rüsten derzeit etwa ein Drittel ihrer Kinosäle mit 3D-Leinwänden aus, da 3D-Filme das Dreifache an Umsätzen erzielen. Marktanalyst *Screen Digest* rechnet bis zum Jahr 2009 weltweit mit 5000 3D-fähigen Kinosälen.

Auch der 3D-Display-Markt zeigt mit wachsenden Umsätzen und neuen Technologien einen deutlichen Aufwärtstrend. Firmen wie Philips mit den entwickelten 42“ 3D-WOW-Displays, LG, Samsung mit der neusten DLP[1]-Technik von Texas Instruments (bis zu 70“ Bildschirmdiagonale) oder Miracube setzten die ersten ernst zu nehmenden Zeichen in Richtung 3D-Home-Entertainment.

Der aktuelle Fortschritt in der 3D-Videokommunikation wird zukünftig den Einsatz neuer Anwendungen und Systeme wie *3D-TV*, *Free-Viewpoint-Video* oder *immersive Videokonferenzsysteme* ermöglichen [161]. Ferner erschließt das Breitbandinternet neue Wege zur Verbreitung medialer Inhalte. Heutzutage ist es schon selbstverständlich, dass Fernsehsendungen der öffentlich-rechtlichen und privaten Rundfunkanstalten über das *World Wide Web* (WWW) jederzeit erreichbar sind. In den kommenden Jahren werden weitere unabhängige Anbieter hinzukommen, die ausschließlich über das WWW senden (z.B. Joost). Diese IPTV-Anwendungen werden die Sehgewohnheiten der jüngeren Generation radikal verändern. *3D-TV over IP* [176] wird mit Sicherheit einen Teil dieser Angebote darstellen und dies umso stärker, je besser und preiswerter die 3D-Displaytechnik für Heimanwendungen wird.

[1] Direct Light Processing

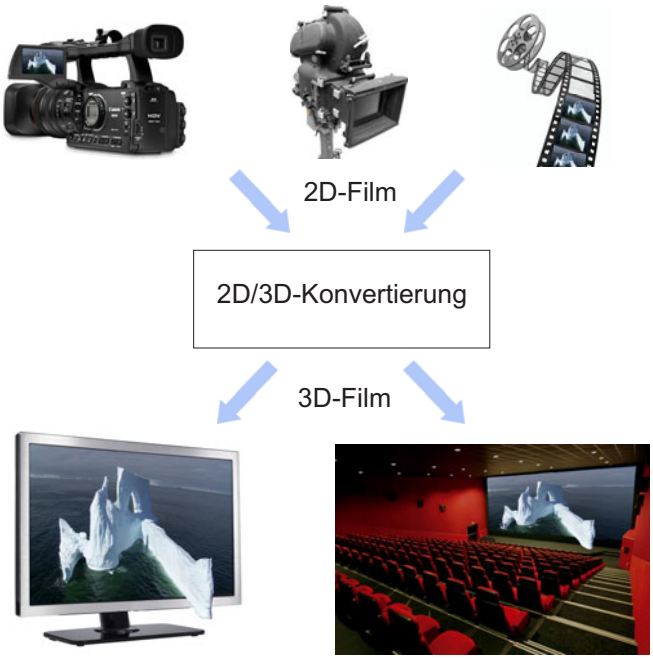


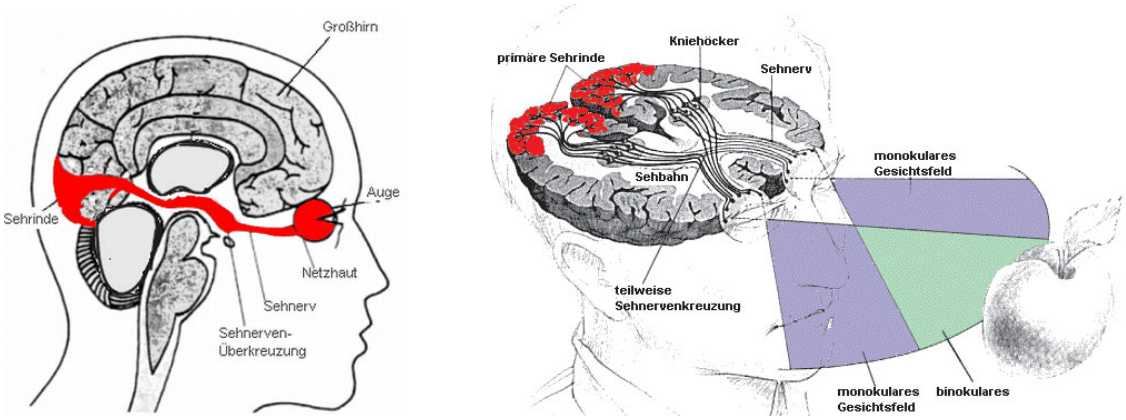
Abbildung 1.1.: 2D/3D-Konvertierung von Filmen

Das größte Problem dieser Entwicklung stellt der geringe Umfang an 3D-Inhalten wie z. B. 3D-Filmen dar, wodurch große Marktbarrieren der neuen 3D-Technologie existieren. Seit einigen Jahren wird ein erheblicher Forschungsaufwand in die Konvertierung zweidimensionaler Videoaufnahmen in dreidimensionale Modelle oder stereoskopische Videoaufnahmen betrieben, um diese Lücke zu schließen. Auf der anderen Seite stößt die Konvertierung von 2D-Filmen in 3D vor allem bei Filmstudios und Rechteinhabern auf großes Interesse. Bereits produzierte Filme mit großen Erfolgen und Einspielergebnissen wie „Star Wars“, „Titanic“, „Zurück in die Zukunft“, etc. könnten als Neuauflage in 3D noch einmal erscheinen und viele Besucher in die 3D-Kinos locken.

Die qualitativ hochwertige Konvertierung eines Films von 2D in 3D ist derzeit nur mit einem enormen manuellen Aufwand möglich und somit kann die zukünftige Nachfrage nach 3D-Inhalten in keinster Weise befriedigt werden. Eine Automatisierung der Konvertierungsprozesse könnte die Kosten und die Produktionszeiten stark reduzieren und somit den Weg ebenen, die 3D-Technologie weiter voranzutreiben.

Von der Motiviation getrieben, alte Filme als Neuauflage in 3D zu erleben, beschäftigt sich die vorliegende Arbeit mit der automatischen Konvertierung von Videosequenzen von 2D in 3D (siehe Abbildung 1.1). Zur Umsetzung dieses Ziels ist der Einsatz und die Weiterentwicklung von Verfahren des *maschinellen Sehens* (engl. *Computer Vision*, CV) und der *Computergrafik* (engl. *Computer Graphics*, CG) notwendig, die in dieser Arbeit beschrieben werden.

Im folgenden Abschnitt wird eine kurze Einführung in die Physiologie des dreidimensionalen Sehens und in die Geschichte der Stereoskopie gegeben.



(a) (b)

Abbildung 1.2.: Gesichtsfeld, Sehbahn und visueller Kortex: (a) Sehbahn zum visuellen Kortex; (b) Gesichtsfeld und Querschnitt durch den Kortex. (Quelle: www.scheffel.og.bw.schule.de)

1.2. Einführung in die Stereoskopie

Bei der Aufnahme einer dreidimensionalen Szene mit einer normalen Film- oder Videokamera gehen durch die Abbildung in der Kamera die Tiefeninformationen verloren. Der Mensch besitzt von Natur aus zwei Augen, die seine Umgebung von zwei leicht abweichenden Perspektiven wahrnehmen. Das Gehirn verarbeitet diese Informationen und generiert daraus einen räumlichen Tiefeneindruck. Um beide Perspektiven zu simulieren, werden dem Betrachter von 3D-Darstellungen zwei unterschiedliche Ansichten einer Szene vorgeführt, jeweils eine pro Auge. Aus diesen stereoskopischen Informationen fügt der visuelle Cortex des Gehirns schließlich eine räumliche Szene zusammen.

1.2.1. Physiologie des 3-D Sehens

Abbildung 1.2(a) zeigt eine seitliche Ansicht des Weges von der Aufnahme optischer Reize im Auge (lat. *Oculus*) über die zentrale Sehbahn (lat. *Nervus opticus*) in das visuelle Zentrum des menschlichen Gehirns, der primären Sehrinde (visueller Kortex), wo ein Abbild der Netzhaut verzerrt dargestellt und verarbeitet wird [98]. Die Sehbahnen des linken und des rechten Auges kreuzen sich dabei teilweise in der Sehnervenkreuzung (lat. *Chiasma opticum*), so dass das linke monokulare Gesichtsfeld im rechten Teil des visuellen Kortex und das rechte monokulare Gesichtsfeld im linken Teil des visuellen Kortex abgebildet werden (siehe Abbildung 1.2(b)). Bei dem binokularen Gesichtsfeld, das für das räumliche Sehen verantwortlich ist, bleiben die Fasern der Sehbahn auf der gleichen Seite der Gehirnhälfte [166].

Durch Fusion der beiden perspektivisch verschiedenen Bilder des binokularen Gesichtsfelds wird ein räumliches Bild wahrgenommen. Abbildung 1.3 zeigt ein Beispiel

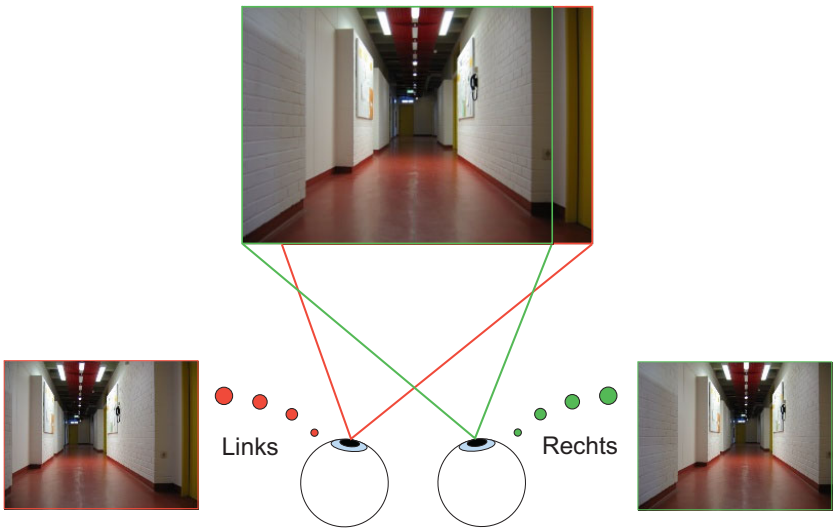


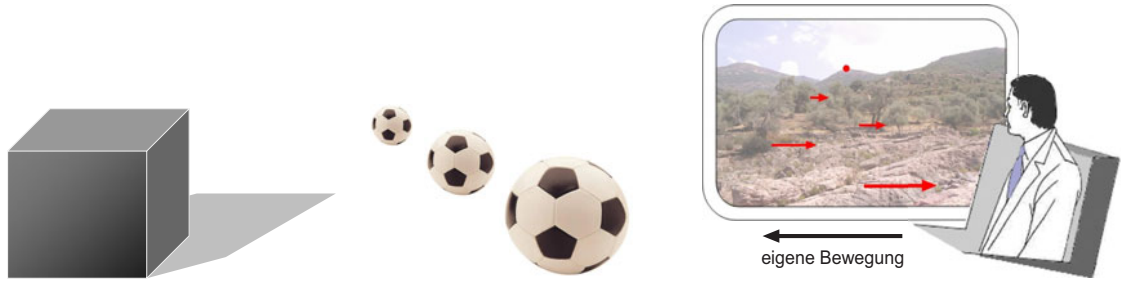
Abbildung 1.3.: Binokulares Sehen: Linkes und rechtes Bild wird jeweils getrennt betrachtet und im Gehirn zusammengefügt.

zweier Bilder, die getrennt voneinander aus unterschiedlicher Perspektive aufgenommen wurden. Die biologische Funktion des räumlichen Sehens ist die genaue Entfernungsmessung und Orientierung in der näheren Umgebung, wie sie z. B. bei Raubtieren zum Jagen (Katzen, Greifvögel, etc.) oder Primaten beim Klettern erforderlich ist. Der Nachteil ist eine größere Einschränkung des Gesichtsfeldes gegenüber Lebewesen, die einen „Rundumblick“ besitzen, um vor allem Gefahren optisch wahrnehmen zu können.

Das binokulare bzw. stereoskopische Sehen stellt allerdings nur einen Teil der Wahrnehmung von Tiefe dar. Das Gehirn ist in der Lage, durch monokulare bzw. extrastereoskopische Informationen, Tiefe zu interpretieren (monokulare Tiefenwahrnehmung). Diese Art der Tiefenwahrnehmung beruht auf Erfahrungswerten. Zu ihr zählen die Bewegungsparalaxe, Licht- und Schatteneffekte, relative Größenunterschiede, Interposition, Texturgradient, Perspektive, Luftperspektive und Helligkeitsanpassung [108].

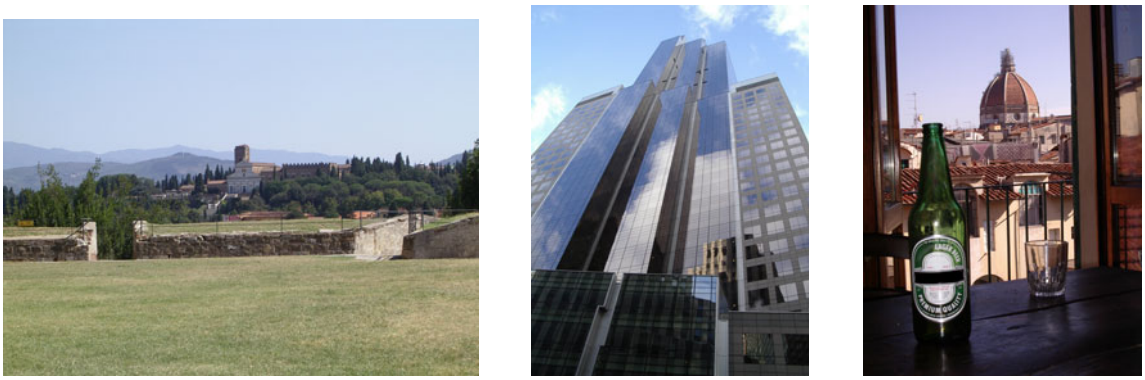
Licht- und Schatten: Abbildung 1.4(a) zeigt ein Beispiel für die Interpretation von Tiefe durch Licht- und Schatteneffekte. Dreidimensionale Objekte werfen je nach Beleuchtungsrichtung Schatten. Zudem erscheinen direkt beleuchtete Bereiche dieser Objekte heller.

Relative Größenunterschiede: Objekte, deren Größe erfahrungsmäßig bekannt ist bzw. die mehrfach in unterschiedlicher Größe sichtbar sind, weisen auf verschiedene Tiefen hin. Die Bälle in Abbildung 1.4(b) besitzen die gleiche absolute Größe aber unterschiedliche relative Größe. Durch relative Größenunterschiede lassen sich ebenfalls Entfernungen schätzen.



(a) (b) (c)

Abbildung 1.4.: Monokulare Tiefenwahrnehmung I: (a) Beleuchtung und Schatten, (b) relative Größenunterschiede und (c) Bewegungsparallaxe.



(a) (b) (c)

Abbildung 1.5.: Monokulare Tiefenwahrnehmung II: (a) Luftperspektive, (b) Perspektive und (c) Interposition.

Bewegungsparallaxe: Die Bewegungsparallaxe hat eine große Bedeutung für die Tiefenwahrnehmung, dies um so stärker je dichter die Objekte sind. Das beste Beispiel für Bewegungsparallaxe lässt sich bei dem Blick aus dem Fenster eines fahrenden Zuges beobachten (siehe Abbildung 1.4(c)). Objekte im Vordergrund bewegen sich schneller als Objekte im Hintergrund.

Luftperspektive: Ein etwas schwächerer Hinweis auf Tiefe ist die Luftperspektive, die vor allem bei Landschaftsaufnahmen auftritt und von Malern verstärkt in Zeichnungen eingesetzt wird. Weiter entfernte Bereiche verlieren ihre Farbe durch das Auftreten von Dunst in der Atmosphäre (siehe Abbildung 1.5(a), Hügel im Hintergrund).

Perspektive: Die Perspektive ist eng verbunden mit dem binokularen Sehen. Parallele Kanten verlaufen in einer perspektivischen Abbildung scheinbar nicht mehr parrallel, sondern schneiden sich in dem sogenannten Fluchtpunkt, der im Unendlichen liegt (siehe Abbildung 1.5(b)).

(a) (b)

Abbildung 1.6.: Monokulare Tiefenwahrnehmung III: (a) Texturgradient (Quelle: www.juliewiegand.com) und (b) Helligkeitsanpassung (Quelle: www.vis.uni-stuttgart.de)

Interposition: Abbildung 1.5(c) zeigt ein einfaches Beispiel für die Tiefenwahrnehmung durch Interposition. Vordergrundobjekte verdecken Bereiche des Hintergrunds.

Texturgradient: Ein Tiefenhinweis, der durch weitere monokulare Tiefenwahrnehmung wie Luftperspektive unterstützt wird, ist der Texturgradient. Farbstrukturen verlieren mit zunehmender Entfernung ihre Textur und erscheinen zunehmend homogen. Als Beispiel dient das gemalte Sonnenblumenfeld in Abbildung 1.6(a).

Helligkeitsanpassung: Die Helligkeitsanpassung, die auch als *Depth Cueing* bezeichnet wird, ist eine in der Computergrafik häufig eingesetzte Technik, um Tiefe zu simulieren. Dabei verlieren Objekte mit zunehmender Entfernung an Helligkeit (siehe Abbildung 1.6(b)).

Das binokulare Sehen zur Tiefenwahrnehmung wird im Vergleich zum monokularen Sehen häufig überschätzt. Dies wird z. B. bei den Auswirkungen deutlich, wenn einzelne monokulare Tiefenhinweise verloren gehen. Die monokulare Tiefenwahrnehmung wird vor allem bei zunehmender Dunkelheit stark verringert und kann durch das binokulare Sehen nur ansatzweise kompensiert werden.

1.2.2. Geschichte der Stereoskopie

Einer der ersten Pioniere, der sich im Jahre 1832 (kurz nach der Erfindung der Fotografie im Jahr 1826) mit dem Problem des räumlichen Sehens beschäftigte, war der englische Physiker Sir Charles Wheatstone. Er konstruierte ein Gerät, welches zwei Bilder, die perspektivisch leicht versetzt gezeichnet wurden, getrennt in das linke und rechte Auge

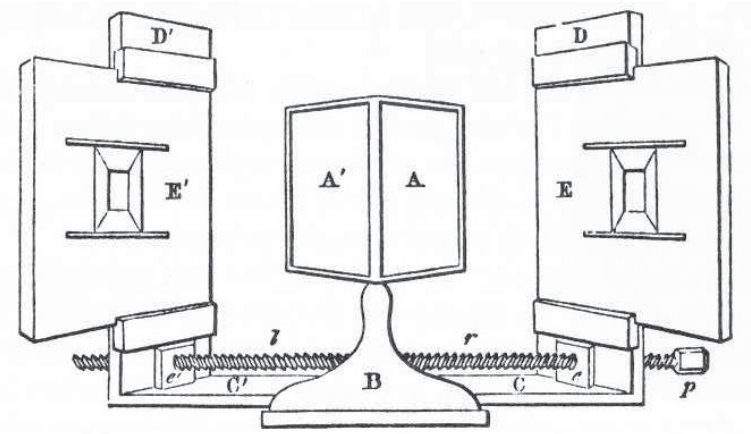


Abbildung 1.7.: Wheatstone's Spiegelstereoskop (Quelle: Carl Wheatstone, 1838).

projizieren konnte und gab dem Gerät den Namen „Stereoskop“[2]. Die Bilder waren auf Holzplatten links und rechts befestigt. In der Mitte wurden zwei senkrecht zueinander liegende Spiegel angebracht, von denen jeweils die Bilder reflektiert wurden. Durch die geringe Entfernung zwischen Betrachter und Spiegel gelangten die Bilder jeweils getrennt voneinander in das linke und rechte Auge und lösten somit einen räumlichen Effekt aus. In Abbildung 1.7 ist eine Zeichnung von Wheatstone's Spiegelstereoskop dargestellt.

Der englische Physiker Sir David Brewster stellte 1849 die erste Fotokamera mit zwei Objektiven vor. Damit konnten zwei Bilder zeitgleich aufgenommen werden, was zuvor nur durch zweimalige Belichtung nacheinander möglich war. Diese Erfindung hatte den großen Vorteil, dass auch bewegliche Motive stereoskopisch fotografiert werden konnten.

Es gibt große Diskussionen darüber, welches die erste Filmproduktion in 3D war. Am häufigsten wird auf den 3D-Film „L'Arrivée d'un train“ der Brüder Louis und Auguste Lumiére, die Erfinder des Kinos, verwiesen, der 1903 gefilmt wurde. Die Geschichte des 3D-Kinos und 3D-Fernsehens lässt sich grob in fünf Epochen unterteilen.

Experimentierphase: Zwischen 1900 und 1943 war die Experimentierphase des 3D-Films, in der einige wenige Filme mit geringem Budget gedreht wurden, um vor allem die Stereoskopie weiter zu erkunden.

1. Goldene Ära des 3D-Films: Zwischen 1950 und 1955 erlebte der 3D-Film mit knapp 100 produzierten Filmen seinen größten Erfolg. Als Ursache für diese Entwicklung wurde vor allem der starke Besucherrückgang in den Kinos gesehen, da das kommerzielle Fernsehen im Heimbereich immer günstiger geworden war. Allerdings endete diese Epoche fast schlagartig nach 1955, wofür mehrere Gründe angenommen werden. Zum einen

[2]Der Begriff „stereo“ stammt aus dem Griechischen und bedeutet soviel wie „körperlich“ bzw. freier übersetzt „räumlich“.

waren die Filme nur von geringem inhaltlichen Niveau mit Qualitätseinbußen durch die schlechten Trenneigenschaften der Rot-Grün-Filtertechnik und zum anderen entwickelte sich das Farbfernsehen vor allem in Nordamerika sehr schnell.

Renaissance: Einige Filmproduktionsstudios versuchten den 3D-Film zwischen 1973 und 1985 neu zu beleben. Allerdings war auch hier die Trennung von rechter und linker Ansicht zu gering, so dass der 3D-Film erneut für kurze Zeit das Interesse der Zuschauer und somit auch der Filmstudios verlor.

Revolution: Zwischen 1986 und 2000 wurde das 3D-Kino durch das neue IMAX 3D Filmformat (70 mm), mit der die Auflösung und die Qualität gegenüber konventionellen Kinoformaten (35 mm) deutlich erhöht wurden, revolutioniert. Allerdings blieb der große Erfolg aufgrund der hohen Produktionskosten von 3D-Filmen und somit einem Mangel an Filmen weitestgehend aus.

2. Goldene Ära: In 2001 startete die 2. goldene Ära des 3D-Films. Die zunehmende Digitalisierung des Films, die Entwicklung computeranimierter Filme aber auch die Entwicklung der 3D-Displaytechnologie gelten als die treibenden Faktoren für die Produktion vieler neuer 3D-Filme. Allerdings wird auch der starke Rückgang der Besucherzahlen in konventionellen Kinos, der vermutlich durch die rasante Entwicklung der Unterhaltungselektronik im Heimbereich (DVD, HD-TV, Dolby Surround) beeinflusst wird, als Grund für diesen 3D-Trend gesehen.

Seit dem Beginn des 3D-Films wurden insgesamt etwa 300 Filme in 3D produziert. Davon sind etwa 70 allein in der 2. goldenen Ära entstanden.

1.3. Stand der Forschung

1.3.1. Tiefenschätzung aus Bildern und Bildfolgen

Die Rückgewinnung der Tiefeninformation aus monokularen Bildern oder Bildfolgen ist ein zentraler Bestandteil in einer Vielzahl von Anwendungsgebieten, weshalb sich dieses Forschungsfeld seit Jahrzehnten großer Beliebtheit erfreut. Viele fundamentale Algorithmen wurden seitdem entwickelt, die aus den Bildinformationen die 3D-Struktur der aufgenommenen Szene rekonstruieren. Diese Methoden lassen sich in zwei grundlegende Ansätze unterteilen: Verfahren, die die 3D-Information ausschließlich aus einzelnen Bildern schätzen (*monokulare Tiefenschätzung*), und Methoden, die auf zwei oder mehreren Ansichten einer Szene basieren.

Monokularen Tiefenschätzverfahren versuchen dabei u.a., die menschlichen Erfahrungswerte bei der monokularen Tiefenwahrnehmung in Algorithmen umzusetzen. Zu diesen Verfahren zählen: Tiefe aus Fokus (engl. *Shape-from-Focus*, SfF) und Defokus (engl. *Shape-from-Defocus*, SfD) [111, 194], Tiefe aus Schattierung (engl. *Shape from Shading*, SfSh) [33, 200] und Tiefe aus Textur (engl. *Shape-from-Texture*, SfT) [15, 69, 70, 73, 86, 105, 137], wobei für das letztere verschiedene Ansätze wie die *strukturelle Texturanalyse* [15, 86], die *statistische Texturanalyse* [69, 73] oder Analyseverfahren, die die geometrischen Szeneneigenschaften wie Fluchtpunkte und Fluchtlinien aus einem Bild extrahieren [70, 105, 137], beinhaltet.

Zu den Verfahren, die auf zwei oder mehreren Ansichten einer Szene basieren, zählen: Tiefe aus Stereo (engl. *Structure from Stereo*, SfS) [3, 5, 35, 37, 92, 94, 156, 205], Tiefe aus Bewegung (engl. *Structure from Motion*, SfM) [2, 9, 24, 84, 85, 144, 171, 174, 179] und Tiefe bzw. Form aus Silhouetten (engl. *Shape-from-Silhouette*, SfSi) [20, 104].

SfS-Verfahren sind Disparitätsschätzverfahren zwischen zwei oder mehreren Ansichten, die von einem festen Kamerasetup zeitgleich aufgenommen werden. Mit ihnen erstellt man Disparitätskarten, die bei bekannter Kamerageometrie in sogenannte Tiefenkarten umgerechnet werden können. Bei SfM-Verfahren wird die Struktur der Szene und die Kamerabewegung (Rotationen, Translationen, interne Kameraparameter) mithilfe einzelner Punkte, die über die Bildsequenz verfolgt werden, berechnet. Hierzu wird lediglich eine einzelne Kamera verwendet, die die Szene von verschiedenen Perspektiven zu unterschiedlichen Zeitpunkten aufnimmt. SfSi-Verfahren hingegen nehmen Objekte meist zeitgleich von unterschiedlichen Perspektiven auf und rekonstruieren diese über die Schnittmenge binärer Segmentierungsmasken im 3D-Raum.

Eine detaillierte Beschreibung dieser Ansätze zur Tiefenschätzung ist in Abschnitt 6.1.2 aufgezeigt.

1.3.2. Synthese virtueller Ansichten

Die Synthese neuer virtueller Ansichten (engl. *Rendering*) kann mit einer Reihe von Methoden durchgeführt werden, die sich in ihren Eigenschaften dahingehend unterscheiden, dass sie verschieden viel Szenengeometrie benötigen. Das aus der Computergrafik bekannte Texture-Mapping [14] benötigt eine möglichst genaue 3D-Geometrie mit Texturinformationen aus einer geringen Anzahl an Bildern. Dem gegenüber steht das *Lightfield Rendering* [106], das aus einer Vielzahl überlappender Bilder neue Ansichten durch Interpolation erzeugt, ohne 3D-Geometrieinformationen zu verwenden. Lightfields basieren auf der Definition der *plenoptischen Funktion* [1], die besagt, dass ein Bild durch die Summe aller Lichtstrahlen, die von Szenenobjekten reflektiert werden, erzeugt wird. Ähnlich dem Lightfield Rendering entstehen beim *Lumigraph* [59] neue Ansichten durch Interpolation,

mit dem Unterschied, dass Bereiche mit geringer Überlappung durch lokale geometrische Approximationen erzeugt werden.

Zwischen diesen Ansätzen stehen Verfahren, die zusätzlich zu der Bildinformation Tiefenkarten verwenden. Zu diesen gehören das sogenannte *Depth Image-based Rendering* (DIBR) [44, 95], *Layered-Depth Images* (LDI) [164] und *Intermediate View Reconstruction* (IVR) [120, 206]. Existiert für jedes Pixel eines Bildes die Tiefeninformation, können neue Ansichten durch entsprechenden Versatz der Pixel generiert werden. LDIs besitzen zudem für jedes Pixel mehrere Farb- und Tiefenwerte, so dass Aufdeckungen durch die verschiedenen Layer kompensiert werden. Bei IVR werden virtuelle Ansichten aus zwei oder mehr benachbarten Ansichten, zu denen jeweils eine Tiefenkarte existiert, erzeugt, wodurch Verdeckungen in der Regel nicht existieren.

Speziell für die Synthese stereoskopischer Ansichten aus monokularen Videosequenzen existieren noch weitere Ansätze, die ohne 3D-Geometrieinformation auskommen. Zu ihnen zählen Verfahren, die z. B. eine räumliche und zeitliche Interpolation bewegter Bilder durchführen [55] bzw. eine planare Transformation auf zeitlich benachbarte Bilder anwenden [152], um somit ein Stereokamerasetup zu simulieren. Allerdings ist bei diesen Ansätzen die zeitliche Konsistenz der Parallaxe über die Bildsequenz nicht gegeben und hängt stark von der 3D-Szene und der Kamerabewegung ab, da ein Stereokamerasetup nicht korrekt modelliert wird. In [31] wird ein Verfahren vorgestellt, welches über eine 2D-Delaunay-Triangulation[3] Tiefen zwischen einzelnen Bildpunkten interpoliert.

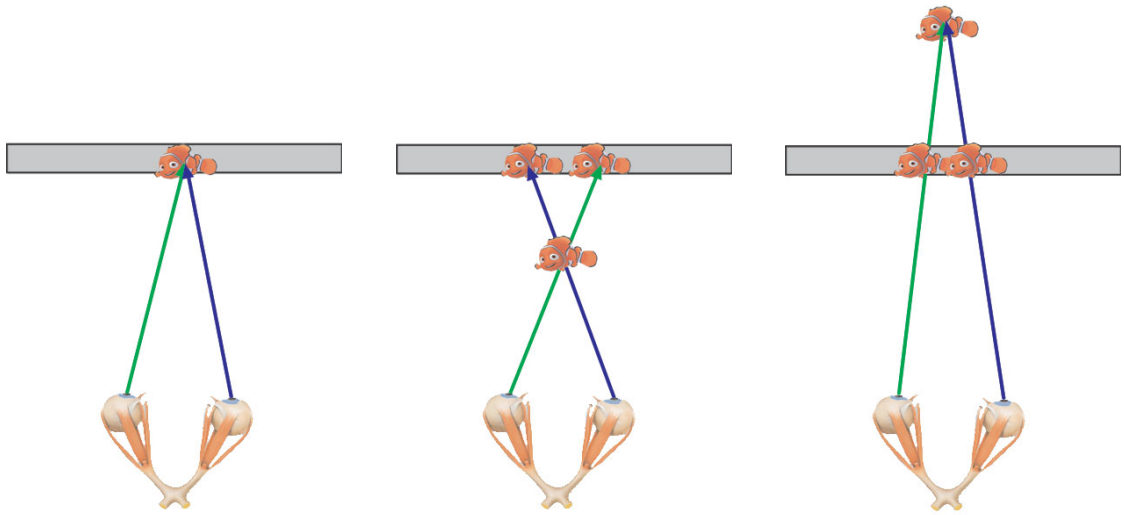
1.3.3. Darstellungsmethoden von stereoskopischen Ansichten

Zur Erzeugung des räumlichen Tiefeneindrucks müssen stereoskopische Ansichten voneinander getrennt in die beiden Augen projiziert werden. Der räumliche Effekt entsteht durch die unterschiedlichen Parallaxen der Objekte in der Szene entsprechend ihrer Tiefe. Da die Stereoansichten auf einer Ebene dargestellt werden, erscheinen die Objekte je nach Art der Parallaxe vor, in oder hinter der Ebene. In Abbildung 1.8 sind die drei Arten der Parallaxe dargestellt.

Ist keine Parallaxe vorhanden (ein Objekt befindet sich exakt an der gleichen Position in beiden Bildern) so wird der Eindruck vermittelt, dass sich das Szenenobjekt unmittelbar in der Bildschirmebene befindet. Bei negativer Parallaxe (ein Objekt ist in der rechten Ansicht weiter links positioniert als in der linken Ansicht) erscheint das Objekt vor der Bildebene. Eine positive Parallaxe (ein Objekt ist in der rechten Ansicht weiter rechts positioniert als in der linken Ansicht) führt dazu, dass Objekte hinter der Ebene liegen.

Da das Gehirn sowohl monokulare als auch binokulare Tiefenhinweise gleichzeitig verarbeitet, müssen alle Tiefenhinweise jederzeit miteinander übereinstimmen, da sonst

[3] Tiefeninterpolation zwischen weiter entfernten Pixeln im Bild



(a) (b) (c)

Abbildung 1.8.: Arten der Parallaxe: (a) ohne Parallaxe, (b) negative Parallaxe und (c) positive Parallaxe.

Beeinträchtigungen der Tiefenwahrnehmung auftreten. In Abbildung 1.9 sind zwei Beispiele gezeigt, bei denen einmal die Tiefenhinweise im Einklang sind (Abbildung 1.9(a)) und einmal im Widerspruch stehen (Abbildung 1.9(b)). Besitzt ein Objekt eine negative Parallaxe, wird aber teilweise durch den Rahmen des Displays verdeckt, stehen monokulare (Interposition) und binokulare Tiefenwahrnehmung im Widerspruch. Demnach sollten sich Objekte, die seitlich aus dem Bildschirm verschwinden, immer hinter der Bildschirmebene befinden (positive Parallaxe).

Für die Trennung der beiden stereoskopischen Ansichten existieren bereits eine Reihe von Techniken. Die wohl bekannteste Darstellungsmethode ist die Anaglyphentechnik, bei der z. B. bei der linken Ansicht nur der rote Kanal und bei der rechten Ansicht nur der grüne und blaue Kanal der RGB-Bilder dargestellt wird. Mit Hilfe einer Rot-Cyan-Brille werden jeweils nur die entsprechenden Kanäle durchgelassen. Ähnliche Anaglyphen wie Rot-Blau, Rot-Grün oder Blau-Gelb sind ebenfalls weit verbreitet. Bei diesem Verfahren gehen allerdings Farbinformationen verloren. Zudem sind die Trenneigenschaften bekanntermaßen schlecht, so dass Nebensprechen der Kanäle entsteht.

Eine sehr alte Technik basiert auf dem Pulfrich-Effekt, der 1922 vom deutschen Physiker Carl Pulfrich entdeckt wurde. Sie beruht im Prinzip auf einer optische Täuschung. Durch die Abdunklung eines Auges gegenüber dem zweiten Auge werden optische Reize vom Gehirn später verarbeitet bzw. wahrgenommen. Dieser Effekt ist allerdings nur bei bewegten Aufnahmen ausnutzbar.

Aus den Anfangszeiten der IMAX-3D Kinos ist die sogenannte Shutter-Technik ebenfalls sehr bekannt, bei der die beiden stereoskopischen Ansichten zeitlich versetzt dargestellt werden. Die Trennung erfolgt durch Synchronisation mit einer Shutter-Brille, deren

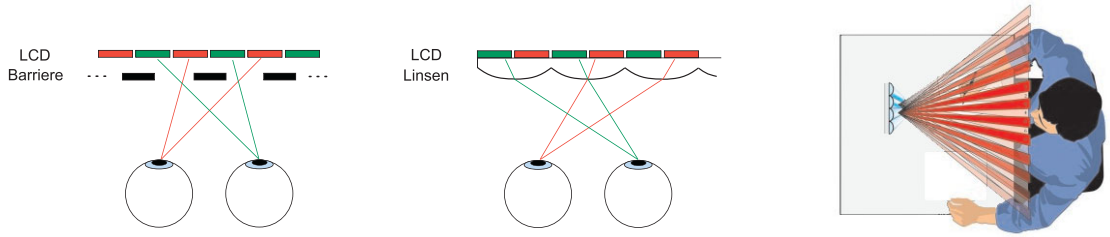
(a) (b)

Abbildung 1.9.: Widerspruch zwischen monokularer und stereoskopischer Tiefenwahrnehmung: (a) Tiefenwahrnehmung im Einklang und (b) Tiefenwahrnehmung gestört.

Gläser aus zwei Flüssigkristallanzeigen bestehen, die wechselnd durchlässig und undurchlässig geschaltet werden. Allerdings besitzt diese Technik neben dem zeitlichen Auflösungsverlust auch einen Helligkeitsverlust durch Polarisation.

Eine weitere Methode, die derzeit in den 3D-Kinos zum Einsatz kommt und die bisher besten Trenneigenschaften aufweist, benötigt ebenfalls eine Brille für die Trennung. Diese Technik verwendet ausschließlich die Polarisationstechnik, d.h. das Licht der linken und der rechten Ansicht wird um 90° gedreht polarisiert (lineare Polarisation). Die Brille besitzt zwei Polarisationsfilter, die lediglich das Licht in der entsprechend richtigen Schwingungsebene durchlässt. Eine Erweiterung der linearen Polarisation ist die zirkulare Polarisation, bei der die beiden Schwingungsebenen mit einer Phasendifferenz rotieren. In einer zirkularen Polarisationsbrille wird erst die Rotation wieder rückgängig gemacht und anschließend das Licht entsprechend der Schwingungsebene durchgelassen. Die zirkulare Polarisation besitzt einen großen Vorteil gegenüber der rein linearen Polarisation, da seitliche Kopfbewegungen kein zusätzliches Nebensprechen verursacht. Allerdings entstehen bei beiden Techniken Helligkeitsverluste durch die Polarisation.

In der 3D-Displaytechnik haben sich zwei Verfahren durchgesetzt, bei denen keine zusätzlichen Brillen zum Einsatz kommen, weshalb diese Displays auch *autostereoskopische* Displays genannt werden. Abbildung 1.10 zeigt das Prinzip dieser beiden Darstellungsmethoden: Parallax-Barriere- und Linsen-Display. Bei dem Parallax-Barriere-Display wird zusätzlich eine Flüssigkristallschicht auf das TFT-LCD aufgebracht, so dass das Licht nebeneinander liegender Pixel in unterschiedlicher Richtung aus dem Bildschirm gestrahlt wird. Werden die Stereobilder vertikal verschachtelt dargestellt und sitzt der Betrachter in der richtigen Position vor dem Bildschirm, werden die beiden Ansichten getrennt in das linke und rechte Auge projiziert. Nachteilig bei diesem Verfahren ist ebenfalls der Helligkeitsverlust durch die zusätzliche LC-Schicht.



(a) (b)

Abbildung 1.10.: Autostereoskopische Darstellungsmethoden: (a) Parallax-Barriere- und Linsen-Display, (b) Linsen-Display (Quelle: www.philips.com/3dsolutions)

Vom Prinzip her sehr ähnlich funktioniert das Linsen-Display, bei dem - im Falle nur eines Betrachters - eine Schicht vertikaler Linsen auf dem TFT-LCD aufgebracht wird, so dass spaltenweise das linke und rechte Stereobild unterschiedlich in die entsprechenden Augen projiziert wird. Nachteilig bei diesen beiden Verfahren ist die Tatsache, dass der Betrachter die richtige Position einnehmen muss, um Nebensprechen oder sogar Pseudo-3D[4] zu verhindern. Abhilfe schafft man sich z. B. mit zusätzlichen Kameras, die die Kopf- bzw. Augenpositionen verfolgen und entweder das Linsen-Panel oder die Stereobilder nachführt [18]. Bei Multi-user Displays werden mehr als zwei Ansichten (i.d.R. 5 bis 9 Ansichten) gleichzeitig dargestellt und mit einem entsprechend angepassten Linsenraster in den Raum projiziert [103]. Ein Betrachter sieht je nach Betrachtungswinkel nur zwei dieser Ansichten.

Ein guter Überblick über die 3D-Displaytechnologie ist in [12] zu finden.

1.4. Überblick über diese Arbeit

Diese Arbeit beinhaltet neue Ansätze zur Konvertierung monokularer Videosequenzen in stereoskopische Sequenzen und deren Erweiterung zur Super-Resolution sowie eine Reihe von Modifikationen bewährter Algorithmen aus der Computer Vision. Ferner wird ein neues Verfahren zur 3D-Rekonstruktion dynamischer Szenen vorgestellt.

1.4.1. Eigene Ansätze

Der Ansatz, den ich in dieser Arbeit verfolge, lässt sich zwischen geometriebasierten Verfahren und DIBR einordnen. Aus der Vielzahl möglicher Verfahren zur Tiefenschätzung habe ich mich für die Rückgewinnung der 3D-Informationen aus der Kamerabewegung entschieden, da SfM allgemein für seine Zuverlässigkeit bekannt ist [168]. Zudem wird die Szene von mehreren Perspektiven aufgenommen, wodurch Synergien mit direkten

[4] Linke und rechte Ansicht sind jeweils vertauscht.

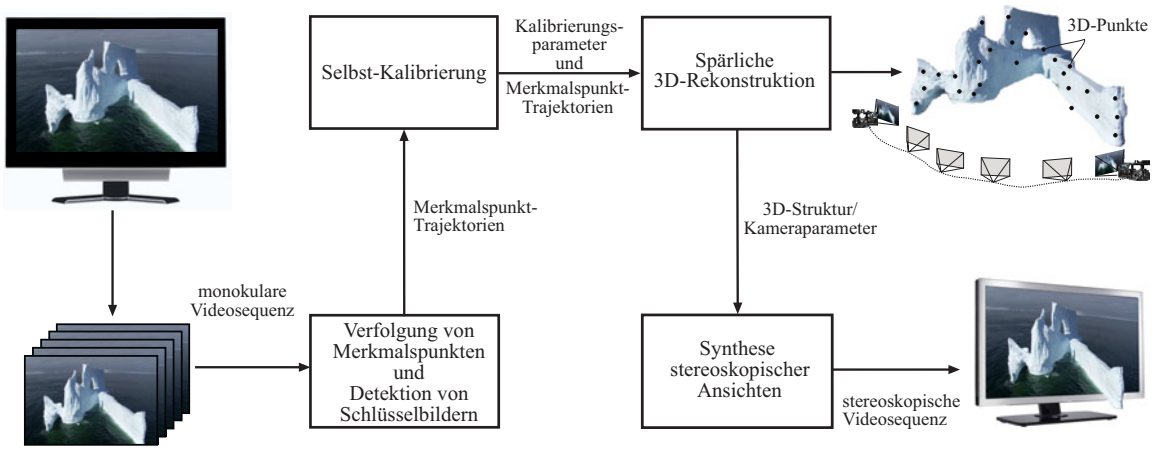


Abbildung 1.11.: Überblick über das 2D/3D-Konvertierungssystem.

Stereoaufnahmen vorhanden sind und damit einige Probleme wie Verdeckungen und Aufdeckungen, die bei der Synthese virtueller stereoskopischer Ansichten in der Regel auftreten, besser gehandhabt werden können.

Bei der Bildsynthese habe ich mich auf bildbasierte Rendering-Methoden konzentriert, da diese fotokonsistente Ergebnisse liefern [123]. Das Ziel dieser Arbeit ist nicht die Rekonstruktion eines kompletten 3D-Modells, sondern die Erzeugung qualitativ hochwertiger stereoskopischer Ansichten, durch die lediglich der räumliche Tiefeneindruck vermittelt werden soll, d.h. das Ergebnis sind zwei perspektivisch versetzte 2D-Ansichten, wobei die zweite Ansicht über den Umweg der Tiefenschätzung virtuell generiert wird. In Abbildung 1.11 ist ein Überblick über das gesamte 2D/3D-Konvertierungssystem dargestellt, dem ich den Namen *Realistic Stereoview Synthesis* (engl., *realistische Synthese stereoskopischer Ansichten*, RSVS) gegeben habe.

Durch die Kamerabewegung entstehen zusätzliche Vorteile, die die Bildqualität erhöhen können. Bei der Aufnahme eines Bildes wird kameraspezifisch eine Abtastung durchgeführt. Durch die Bewegung der Kamera und damit durch eine erneute Abtastung von einer leicht versetzten Perspektive entstehen Redundanzen, mit denen die Auflösung der Bilder erhöht werden können. In dieser Arbeit habe ich diese Redundanzen für die Super-Resolution ausgenutzt, so dass die Auflösung sowohl für die Originalvideosequenz als auch die virtuelle Stereosequenz erhöht wird.

Der neuartige Konvertierungsansatz setzt somit auf die Kombination der leistungsfähigen Algorithmen von SfM [85] und die Idee des bildbasierten Renderns [123] um fotokonsistente stereoskopische Ansichten aus monokularen Videosequenzen zu erzeugen, ohne auf dichte Tiefenkarten zurückgreifen zu müssen.

1.4.2. Wesentliche Beiträge und neue Ansätze

Die wesentlichen Beiträge und Ergebnisse dieser Arbeit sind:

- ein umfangreicher Überblick über den Stand der Forschung
- eine detaillierte Einführung in grundlegende Ansätze der Computer Vision
- eine umfangreiche Untersuchung verschiedener robuster Schätzverfahren
- eine genaue Untersuchung von Kostenfunktionen bei der Verfolgung von Merkmalspunkten
- eine Erweiterung der Merkmalspunktverfolgung basierend auf den Untersuchungsergebnissen
- eine umfangreiche Untersuchung von linearen und nichtlinearen Optimierungsansätzen sowie von Kostenfunktionen bei der Rekonstruktion dreidimensionaler Szenen
- RSVS, ein neuer Ansatz zur Synthese stereoskopischer und Multiview-Ansichten
- eine Erweiterung von RSVS mittels Super-Resolution
- eine umfangreiche Evaluierung von RSVS im Vergleich zu aktuellen Standardverfahren mit standardkonformen subjektiven Qualitätstests
- ein neues Verfahren zur 3D-Rekonstruktion dynamischer Szenen

1.4.3. Aufbau der Arbeit

In Kapitel 2 werden grundlegende geometrische Zusammenhänge beim Kameramodell sowie der projektiven- und der Epipolargeometrie erläutert, die für das weitere Verständnis dieser Arbeit notwendig sind. Robuste Schätzverfahren, die bei der 3D-Rekonstruktion häufig Verwendung finden, werden in Kapitel 3 vorgestellt und untersucht. Kapitel 4 beschäftigt sich mit grundlegenden aktuellen Verfahren zur 3D-Rekonstruktion aus zwei Ansichten. Dieses Kapitel soll als theoretischer Hintergrund für Kapitel 5 dienen, in dem die 3D-Rekonstruktion aus mehreren Ansichten (z. B. Videosequenzen) beschrieben wird.

Kapitel 6 befasst sich mit der eigentlichen Synthese stereoskopischer Ansichten. Hier wird neben Standardverfahren ein neues Syntheseverfahren vorgestellt und beschrieben, welches auf 2D-Transformationen basiert. Die Evaluierung des Syntheseverfahrens im Vergleich zu Standardansätzen wird in Kapitel 7 mithilfe von subjektiven Qualitätstests durchgeführt.

Kapitel 8 liefert eine Einführung in den Umgang mit dynamischen Szenen, wobei ein geometrisches Segmentierungsverfahren vorgestellt wird, welches bewegte Objekte innerhalb einer statischen Szene bei gleichzeitiger Bewegung der Kamera detektiert. Kapitel 9 schließt diese Arbeit mit einer Zusammenfassung und einem Ausblick ab.

Das 2D/3D-Konvertierungssystem aus Abbildung 1.11 ist in den Kapiteln 5 und 6 beschrieben. Der versierte Leser, der sich mit den grundlegenden geometrischen Zusammenhängen, robusten Schätzverfahren und allgemein mit Ansätzen der 3D-Rekonstruktion auskennt, kann die Kapitel 2 bis 4 überspringen und direkt mit Kapitel 5 beginnen.

2. Theoretische Vorbetrachtung

Dieses Kapitel beschreibt die wichtigsten mathematischen und geometrischen Grundlagen bezüglich der Abbildung einer dreidimensionalen Szene auf einer Ebene. Die Beschreibungen sind für ein klareres Verständnis weiterer Problemstellungen in dieser Arbeit wesentlich. Hierbei werden vor allem die Eigenschaften der projektiven Geometrie erläutert, verschiedene Kameramodelle vorgestellt und letztendlich die Geometrie zwischen zwei Ansichten, die Epipolargeometrie und die Homographie, beschrieben, welche die Grundlage der in dieser Arbeit aufgeführten Ansätze zur Rekonstruktion einer dreidimensionalen Szene bilden. Es wird des Öfteren zwischen der *projektiven* und der *affinen* Koordinatendarstellung, die im folgenden Abschnitt näher erläutert werden, gewechselt, um die Herleitung einiger geometrischer Zusammenhänge besser nachvollziehen zu können. Die theoretischen Grundlagen der zugrunde liegenden Geometrie werden zudem von einem algebraischen Blickwinkel betrachtet, um weitere Eigenschaften aufzuführen.

2.1. Projektive Geometrie

Betrachtet man ein Foto einer aufgenommenen Umgebung, so kann man sehr schnell feststellen, dass z. B. Kreise oder Quadrate nicht mehr in ihren ursprünglichen Formen erscheinen. Je nach Betrachtungswinkel der Szene werden Kreise z. B. elliptisch oder Quadrate trapezförmig in dem Foto abgebildet. Diese Art der Abbildung wird auch als projektive Transformation bezeichnet. Eigenschaften wie Winkel, Abstände oder Längenverhältnisse bleiben bei der projektiven Transformation nicht erhalten. Eine wesentliche Eigenschaft der projektiven Geometrie hingegen ist, dass Geraden weiterhin als Geraden dargestellt werden. Abbildung 2.1 zeigt zwei Beispielbilder zweier Szenen, die mit einem Fotoapparat aufgenommen wurden. In dem linken Foto scheinen sich zwei Schienen, die eigentlich parallel verlaufen, in einem Fluchtpunkt etwa in der Bildmitte zu schneiden. Die rechte Aufnahme zeigt einen Korridor mit Bilderrahmen an den Wänden, bei denen die rechten Winkel und die Längenverhältnisse und Parallelitäten der Seiten durch die Projektion auf die Bildebene nicht erhalten bleiben.

Geht man aber zunächst von der dem menschlichen Verständnis am ehesten entsprechenden Euklidischen Geometrie aus, so wird für die Beschreibung einer dreidimensiona-

(a) Gleise (Quelle: www.connective.biz) (b) Korridor

Abbildung 2.1.: Beispiel projektiver Verzerrungen bei der Darstellung einer 3-dimensionalen Szene in einer Bildebene

len Szene gewöhnlich eine Darstellung im Euklidischen Raum verwendet. Ein Punkt auf einer Ebene in $\mathbb{R}^2$ wird in *inhomogenen* oder *affinen* Koordinaten angegeben mit

$$\tilde{\mathbf{m}} = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2. \tag{2.1}$$

Im projektiven Raum $\mathbb{P}^2$ wird dieser Punkt in *homogenen* oder *projektiven* Koordinaten durch Hinzufügen einer 3. Koordinate, die zusätzlich die Lage der 2-dimensionalen Ebene $\mathbb{R}^2$ in $\mathbb{P}^2$ beschreibt, dargestellt:

$$\mathbf{m} = \begin{pmatrix} x \\ y \\ w \end{pmatrix} \in \mathbb{P}^2, \tag{2.2}$$

wobei w ein Skalierungsfaktor und damit die Projektion des Punktes entlang einer Dimension von $\mathbb{P}^2$ ist. Punkte in inhomogenen Koordinaten werden im weiteren Verlauf dieser Arbeit mit einer Tilde gekennzeichnet.

Der n-dimensionale projektive Raum $\mathbb{P}^n$ ist allgemein definiert über die Gleichheitsbedingung zweier Punkte, die sich auf einer Geraden durch das Projektionszentrum (der Ursprung des Koordinatensystems) befinden, mit:

$$\mathbf{m}_1 = \lambda \mathbf{m}_2 \quad \forall \lambda \in \mathbb{R} \backslash \{0\}. \tag{2.3}$$

Die Beziehung zwischen $\mathbb{P}^n$ und $\mathbb{R}^n$ lässt sich über eine injektive Abbildung wie folgt definieren:

$$\mathbf{m} = (m_1, m_2, \ldots, m_n, w)^T \in \mathbb{P}^2 \mapsto \tilde{\mathbf{m}} = \left(\frac{m_1}{w}, \frac{m_2}{w}, \ldots, \frac{m_n}{w}\right)^T \in \mathbb{R}^2. \tag{2.4}$$

Somit enthält ein Vektor in $\mathbb{P}^n$ insgesamt $(n+1)$ Elemente. Einen Sonderfall stellt der Vektor $(m_1, m_2, \ldots, m_n, 0)^T$ im projektiven Raum $\mathbb{P}^n$ dar, welcher einen Punkt im Unendlichen repräsentiert. Für Punkte, die nicht im Unendlichen liegen, kann zur Vereinfachung weiterer Überlegungen der Skalierungsfaktor $w = 1$ gesetzt werden.

2.1.1. Homogene Darstellung von Punkten und Linien in $\mathbb{P}^2$

Eine Linie in der projektiven Ebene wird über die Geradengleichung $ax + by + c = 0$ dargestellt bzw. vektoriell mit $\mathbf{l} = (a, b, c)^T$. Eine Multiplikation mit einem Skalierungsfaktor λ repräsentiert dieselbe Linie auf der Ebene:

$$\mathbf{l} = (a, b, c)^T = \lambda \cdot (a, b, c)^T \quad \forall \lambda \in \mathbb{R} \backslash \{0\}. \tag{2.5}$$

Ein Punkt $\tilde{\mathbf{m}} = (x, y)^T$ liegt auf der Linie $\mathbf{l} = (a, b, c)^T$, wenn die Gleichung $ax + by + c = 0$ erfüllt ist. Vektoriell wird dies in projektiven Koordinaten mit

$$\mathbf{m}^T \mathbf{l} = (x, y, 1) \begin{pmatrix} a \\ b \\ c \end{pmatrix} = 0 \tag{2.6}$$

ausgedrückt. Entsprechend Gl 2.5 gilt bei Multiplikation des Punktes $\mathbf{m} = (x, y, 1)^T$ mit einem Skalierungsfaktor λ:

$$\lambda \mathbf{m}^T \mathbf{l} = (\lambda x, \lambda y, \lambda) \begin{pmatrix} a \\ b \\ c \end{pmatrix} = 0 \quad \forall \lambda \in \mathbb{R} \backslash \{0\}. \tag{2.7}$$

Der Schnittpunkt zweier Linien $\mathbf{l}_1$ und $\mathbf{l}_2$ wird über das Vektor- bzw. Kreuzprodukt berechnet:

$$\mathbf{m} = \mathbf{l}_1 \times \mathbf{l}_2. \tag{2.8}$$

Der Punkt $\mathbf{m}$ liegt somit auf beiden Linien und erfüllt die Gleichungen $\mathbf{m}^T \mathbf{l}_1 = 0$ und $\mathbf{m}^T \mathbf{l}_2 = 0$. Das Kreuzprodukt wird häufig auch - ebenso in dieser Arbeit - in der Form

$$\mathbf{m} = [\mathbf{l}_1]_\times \, \mathbf{l}_2 \tag{2.9}$$

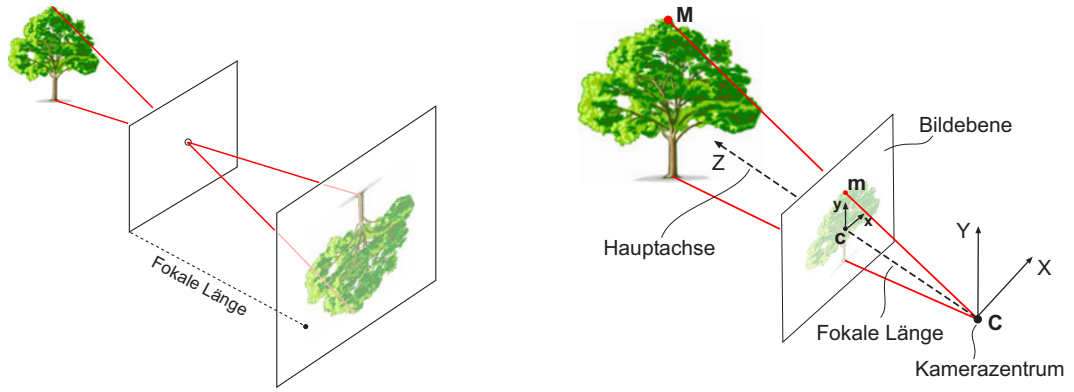
angegeben, wobei $[\mathbf{l}_1]_\times$ eine antisymmetrische 3×3 - Matrix

$$[\mathbf{l}_1]_\times = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix} \tag{2.10}$$

mit den Elementen des Vektors $\mathbf{l}_1$ auf den Nebendiagonalen ist.

2.2. Kameramodell

Die Abbildung einer dreidimensionalen realen Szene auf eine Ebene wird über eine Kamera bzw. Kameramodell beschrieben. In der Literatur sind eine Vielzahl von Kameramodellen beschrieben [39, 68, 160], die unterschiedliche Eigenschaften aufweisen. In den folgenden beiden Abschnitten werden zwei Kameramodelle vorgestellt: die Lochkamera bzw. Zentralprojektion als vereinfachte Darstellung der Projektion und die allgemeine Form einer realen Kamera.



(a) Lochkamera (b) Zentralprojektion

Abbildung 2.2.: Darstellung zweier Kameramodelle: Lochkamera und Zentralprojektion

2.2.1. Lochkamera und Zentralprojektion

Abbildung 2.2 zeigt zwei vereinfachte Darstellungen eines Kameramodells: die Lochkamera und die Zentralprojektion. Beide Modelle besitzen die gleichen Eigenschaften. Da bei dem Lochkameramodell, welches eine spezielle Form der Zentralprojektion ist, die aufgenommene Szene gespiegelt auf der Bildebene dargestellt wird, ist die Zentralprojektion in Abbildung 2.2(b) als Darstellungsform weiter verbreitet.

Liegt das *Projektions-* oder *Kamerazentrum* $\mathbf{C}$ im Ursprung des *Weltkoordinatensystems* in $\mathbb{R}^3$, dann befindet sich die Bildebene, auch *fokale Ebene* genannt, im Abstand $Z = f$ (*fokale Länge*) vom Kamerazentrum. Ein Punkt $\tilde{\mathbf{M}} = (X, Y, Z)^T$ in $\mathbb{R}^3$ wird über die Verbindungslinie mit dem Kamerazentrum auf der Bildebene in Punkt $\tilde{\mathbf{m}} = (fX/Z, fY/Z)^T$ in $\mathbb{R}^2$ abgebildet. In homogenen Koordinaten gilt für die Abbildung des Punktes $\mathbf{M}$ in $\mathbb{P}^3$ auf die Bildebene $\mathbb{P}^2$:

$$\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \longmapsto \begin{pmatrix} fX \\ fY \\ Z \end{pmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}. \tag{2.11}$$

Die Abbildungsmatrix wird auch als *Kameraprojektionsmatrix* oder kurz als *Projektionsmatrix* P bezeichnet, so dass Gl. 2.11 formuliert werden kann mit

$$Z\mathbf{m} = P\mathbf{M}, \tag{2.12}$$

wobei $\mathbf{m} = (fX/Z, fY/Z, 1)^T$ ist. Die 3×4 - Matrix P in Gl. 2.12 stellt die einfachste Form der Projektionsmatrix dar.

2.2.2. Verallgemeinerte Zentralprojektion

Allgemein ist die Projektionsmatrix von vollem Rang und besitzt 11 Freiheitsgrade. Sie kann über QR-Faktorisierung mit

$$\begin{aligned} P &= \begin{bmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \\ &= K\,[R|\mathbf{t}] \end{aligned} \tag{2.13}$$

angegeben werden [83], wobei K die verallgemeinerte Form der *Kamerakalibrierungsmatrix* mit den *internen* oder *intrinsischen* Kameraparametern, R eine Rotationsmatrix und $\mathbf{t}$ ein Translationsvektor ist. R und $\mathbf{t}$ beschreiben die Orientierung und die Position der Kamera bezüglich eines Weltkoordinatensystems und beinhalten die *externen* oder *extrinsischen* Kameraparameter. Die intrinsischen Parameter α_x und α_y repräsentieren die fokale Länge der Kamera in Pixelkoordinaten in x- und y-Richtung, x_0 und y_0 sind die Koordinaten des *Kamerahauptpunktes* (engl. *principal point*), welcher der Schnittpunkt der *Hauptachse* mit der Bildebene ist, und s entspricht dem Scherungsparameter der Pixel des Sensorfeldes, welcher für die meisten realen Kameras näherungsweise Null ist. Ein 3D-Weltpunkt wird über

$$\begin{pmatrix} x_b \\ y_b \\ 1 \end{pmatrix} = K\,[R|\mathbf{t}] \begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix} \tag{2.14}$$

direkt in die Bildkoordinaten $(x_b, y_b)^T$ überführt.

Abbildung 2.3 zeigt ein Beispiel für die verallgemeinerte Zentralprojektion, bei der die Kamera nicht im Ursprung des Weltkoordinatensystems liegt. Unter Verwendung der inhomogenen Koordinaten des Kamerazentrums lässt sich Gl. 2.13 ausdrücken mit

$$P = KR\left[\underline{I}\,|\,-\tilde{\mathbf{C}}\right]. \tag{2.15}$$

$\underline{I}$ ist dabei die 3×3 - Einheitsmatrix.

2.3. Epipolargeometrie

Nachdem die geometrischen und algebraischen Grundlagen der Darstellung einer dreidimensionalen Szene hinsichtlich einer einzelnen projektiven Ansicht bekannt sind, werden in diesem Abschnitt weiterführende Beziehungen und Eigenschaften bezüglich einer zusätzlichen Ansicht der Szene hergeleitet.

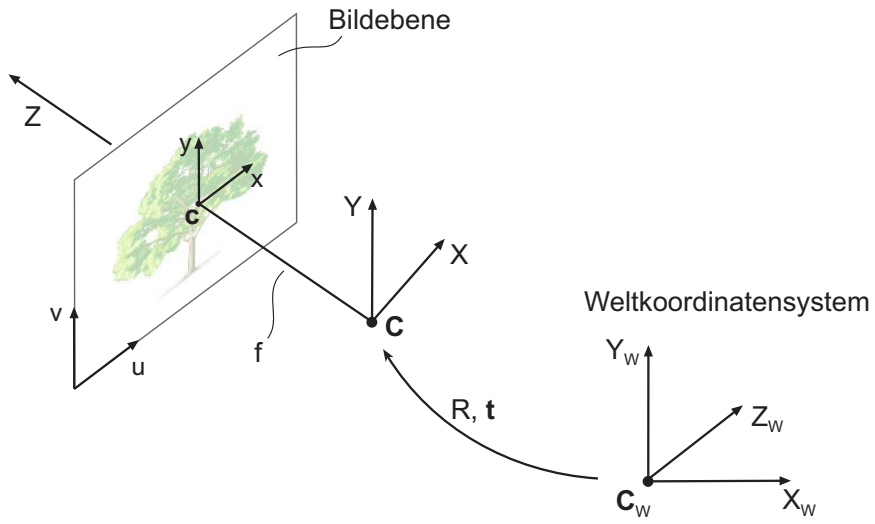


Abbildung 2.3.: Verallgemeinerte Zentralprojektion

Die Beziehungen zwischen zwei Kameraansichten werden durch die Epipolargeometrie - bzw. in algebraischer Form durch die Fundamentalmatrix - beschrieben. Die Epipolargeometrie hängt dabei lediglich von den beiden Kameras (den internen und externen Kameraparametern) und nicht von der 3D-Struktur ab. In Abbildung 2.4 werden die grundlegenden Zusammenhänge der Epipolargeometrie dargestellt. Die *Basislinie*, welche die beiden Kamerazentren verbindet, schneidet die Bildebenen in den *Epipolen*. Zusammen mit dem Strahl, der durch die Rückprojektion des Punktes $\mathbf{m}$ in den 3D-Raum gebildet wird, spannen die beiden Geraden die *epipolare Ebene* auf, welche die Bildebenen in den *Epipolarlinien* schneidet. Demzufolge befindet sich der zu $\mathbf{m}$ korrespondierende Punkt $\mathbf{m}'$, der der Projektion des 3D-Punktes $\mathbf{M}$ in der Bildebene der zweiten Kamera entspricht, auf der Epipolarlinie $\mathbf{l}'$. Diese Eigenschaft wird in vielen Anwendungsszenarien, die sich mit Disparitätsschätzung befassen, ausgenutzt, da der Suchbereich für einen korrespondierenden Bildpunkt bei Kenntnis der Epipolargeometrie von einer 2-dimensionalen Bildebene auf eine 1-dimensionale Linie reduziert werden kann.

2.3.1. Fundamentalmatrix

Algebraisch wird die Epipolargeometrie durch die Fundamentalmatrix repräsentiert, die sich mittels Projektion einfach herleiten lässt:

$$\mathbf{m} = K\left[R|\mathbf{t}\right]\mathbf{M} = K\left(R\tilde{\mathbf{M}} + \mathbf{t}\right) \tag{2.16}$$

$$\mathbf{m}' = K'\left[R'|\mathbf{t}'\right]\mathbf{M} = K'\left(R'\tilde{\mathbf{M}} + \mathbf{t}'\right). \tag{2.17}$$

Der 3D-Weltpunkt $\mathbf{M}$ wird über die Gleichungen 2.16 und 2.17 in die beiden Kameras projeziert. Stellt man Gl. 2.16 nach $\tilde{\mathbf{M}}$ um

$$\tilde{\mathbf{M}} = R^{-1}\left(K^{-1}\mathbf{m} - \mathbf{t}\right) \tag{2.18}$$

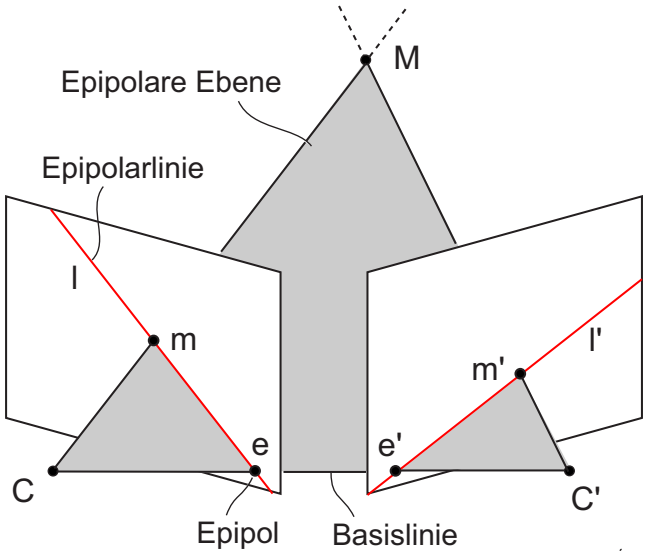


Abbildung 2.4.: Epipolargeometrie: Grundlegende Beziehungen zwischen zwei Kameraansichten

und setzt das Ergebnis in Gl. 2.17 ein, so erhält man

$$\mathbf{m}' = K' \left[R' R^{-1} \left(K^{-1}\mathbf{m} - \mathbf{t} \right) + \mathbf{t}' \right] \tag{2.19}$$

bzw.

$$K'^{-1}\mathbf{m}' = R' R^{-1} K^{-1}\mathbf{m} + \left(\mathbf{t}' - R' R^{-1}\mathbf{t} \right). \tag{2.20}$$

Werden beide Seiten der Gleichung mit $T = \left[\mathbf{t}' - R' R^{-1}\mathbf{t} \right]_\times$ linksseitig multipliziert, ergibt sich

$$TK'^{-1}\mathbf{m}' = TR' R^{-1} K^{-1}\mathbf{m}. \tag{2.21}$$

Durch erneute linksseitige Multiplikation mit $\left(K'^{-1}\mathbf{m}' \right)^T$ wird aus Gl. 2.21

$$\begin{aligned} \left(K'^{-1}\mathbf{m}' \right)^T TK'^{-1}\mathbf{m}' &= \left(K'^{-1}\mathbf{m}' \right)^T TR' R^{-1} K^{-1}\mathbf{m} \\ &= \mathbf{m}'^T K'^{-1} TR' R^{-1} K^{-1}\mathbf{m}. \end{aligned} \tag{2.22}$$

Da das Kreuzprodukt $T\left(K'^{-1}\mathbf{m}' \right)$ ein Vektor ist, der senkrecht auf der von den beiden Vektoren aufgespannten Ebene steht, muss das Skalarprodukt $\left(K'^{-1}\mathbf{m}' \right)^T TK'^{-1}\mathbf{m}' = 0$ sein. Demzufolge ist auch die rechte Seite von Gl. 2.22 gleich Null:

$$\mathbf{m}'^T K'^{-1} TR' R^{-1} K^{-1}\mathbf{m} = 0. \tag{2.23}$$

Mit $F = K'^{-1} TR' R^{-1} K^{-1}$ erhält man die *Epipolargleichung* bzw. *Epipolarbedingung*

$$\mathbf{m}'^T F\mathbf{m} = 0, \tag{2.24}$$

wobei F die 3×3 - Fundamentalmatrix ist, die sich nur aus den internen und externen Kameraparametern zusammensetzt.

Die Epipolarlinie $\mathbf{l}'$, auf der sich der Punkt $\mathbf{m}'$ im zweiten Bild befindet ist

$$\mathbf{l}' = F\mathbf{m}. \tag{2.25}$$

Somit wird aus Gl 2.24

$$\mathbf{m}'^T F\mathbf{m} = \mathbf{m}'^T\mathbf{l}' = 0. \tag{2.26}$$

2.3.2. Essentialmatrix

Sind die internen Parameter der beiden Kameras bekannt, lässt sich die Epipolargleichung (Gl. 2.24) in der Form

$$\mathbf{m}'^T K'^{-1} E K^{-1}\mathbf{m} = 0 \tag{2.27}$$

angeben. $E = TR'R^{-1}$ wird als *Essentialmatrix* bezeichnet und beinhaltet lediglich die externen Kameraparameter. Die Beziehung zwischen Fundamental- und Essentialmatrix lautet demnach

$$F = K'^{-1} E K^{-1}. \tag{2.28}$$

Befindet sich das Weltkoordinatensystem im Ursprung der ersten Kamera, lässt sich Gl. 2.23 mit $R = \underline{I}$ und $\mathbf{t} = \mathbf{0}$ vereinfachen zu

$$\mathbf{m}'^T K'^{-1} \left[\mathbf{t}'\right]_\times R' K^{-1}\mathbf{m} = 0. \tag{2.29}$$

Damit ist $E = \left[\mathbf{t}'\right]_\times R'$.

2.3.3. Eigenschaften der Fundamental- und Essentialmatrix

Die Essentialmatrix, welche sich ausschließlich aus den externen Parametern der Kameras zusammensetzt, besitzt insgesamt fünf Freiheitsgrade - drei für die Rotation und zwei für die Translation - und kann nur bis zu einem Skalierungsfaktor bestimmt werden [160]. Da die Determinate einer antisymmetrischen Matrix Null ist [57], gilt:

$$\mathbf{det}(E) = \mathbf{det}\left(\left[t\right]_\times\right)\ \mathbf{det}(R) = 0. \tag{2.30}$$

Somit ist die Matrix E nicht von vollem Rang (E hat den Rang 2) und besitzt nur zwei unabhängige Spalten- oder Zeilenvektoren.

Für die Fundamentalmatrix können die gleichen Überlegungen angestellt werden. Da F nur bis zu einem Skalierungsfaktor bestimmt werden kann und die Determinante aufgrund von $\mathbf{det}(E) = 0$ ebenfalls Null sein muss, besitzt sie insgesamt sieben unabhängige Parameter. Für die Schätzung dieser Parameter werden somit mindestens sieben Punktkorrespondenzen benötigt. In Kapitel 3 werden Ansätze zur Schätzung der Fundamentalmatrix behandelt.

2.4. Homographie zwischen zwei Ansichten

Ist die Basislänge zwischen zwei Kameras gleich null, d.h. beide Kameras besitzen das gleiche Kamerazentrum, und sei das Kamerazentrum der Ursprung des Weltkoordinatensystems, lässt sich mit $R = \underline{I}$ Gl. 2.19 weiter vereinfachen zu

$$\mathbf{m}' = K'R'K^{-1}\mathbf{m} = H\mathbf{m}. \tag{2.31}$$

Die Matrix H ist eine 3×3 - Matrix und wird als *Homographie* bezeichnet:

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}. \tag{2.32}$$

Nach Gl. 2.31 beschreiben Homograpien eine 2D-Transformation von Punkten einer Ebene auf eine andere Ebene. Geometrisch lässt sich die 2D-Transformation wie folgt deuten: Bei der Epipolargeometrie ist die Epipolarlinie, auf der sich ein korrespondierender Punkt zwischen zwei Ansichten befindet, die Abbildung des in den 3D-Raum rückprojezierten Strahls in der jeweils anderen Ansicht. Sind die Kamerazentren der beiden Ansichten identisch, so ist die Abbildung des rückprojezierten Strahls jeweils lediglich ein Punkt. Somit lassen sich Punkte zweier Ebenen über die Homographie eindeutig zuordnen.

Weiterführende Überlegungen, für die eine eindeutige Zuordnungen von Punkten zwischen Ebenen gilt, sind nach Gl. 2.31, dass nur die internen Parameter der Kameras (z. B. Zoom) variieren. Ferner können bei beliebiger Kamerabewegung (unterschiedliche Kamerazentren) Punkte, die im Unendlichen liegen, bzw. Punkte die auf einer Ebene in $\mathbb{R}^3$ liegen, über eine Homographie eindeutig zugeordnet werden [39].

Beispiel: Als Beispiel kann die Projektion des Mondes, welcher näherungsweise im Unendlichen liegt, auf die Bildebenen zweier Kameras mit unterschiedlichen Kamerazentren und Orientierungen herangezogen werden. Sämtliche Punkte auf dem abgebildeten Mond in der einen Bildebene lassen sich über eine Homographie in die Bildebene der anderen Kamera transformieren.

Mit $\mathbf{m} = (x, y, 1)^T$ und $\mathbf{m}' = \lambda(x', y', 1)^T$ [1] lässt sich Gl. 2.31 komponentenweise notieren:

$$\begin{aligned} \lambda x' &= h_{11}x + h_{12}y + h_{13} \\ \lambda y' &= h_{21}x + h_{22}y + h_{23} \\ \lambda &= h_{33}\left(\frac{h_{31}}{h_{33}}x + \frac{h_{32}}{h_{33}}y + 1\right). \end{aligned}$$

[1] Der Vorfaktor λ wurde hier entsprechend Gl. 2.3 wieder eingeführt, um die Allgemeingültigkeit des projektiven Raums zu verdeutlichen.

Durch Umstellen des Gleichungssystems erhält man die beiden Gleichungen der *perspektivischen Transformation*, mit denen die Koordinaten eines über die Homographie H transformierten Punktes in x- und y-Richtung ermittelt werden können:

$$x' = \frac{a_0 + a_1 x + a_2 y}{1 + c_1 x + c_2 y} \tag{2.33}$$

$$y' = \frac{b_0 + a_1 x + b_2 y}{1 + c_1 x + c_2 y} \tag{2.34}$$

$$\begin{aligned} \text{mit} \quad a_0 &= \frac{h_{13}}{h_{33}},\ a_1 = \frac{h_{11}}{h_{33}},\ a_2 = \frac{h_{12}}{h_{33}}, \\ b_0 &= \frac{h_{23}}{h_{33}},\ b_1 = \frac{h_{21}}{h_{33}},\ b_2 = \frac{h_{22}}{h_{33}}, \\ c_1 &= \frac{h_{31}}{h_{33}},\ c_2 = \frac{h_{32}}{h_{33}}. \end{aligned}$$

Die perspektivische Transformation besitzt insgesamt acht unabhängige Parameter, welche mit genau vier Punktkorrespondenzen, die jeweils zwei Gleichungen (Gl. 2.33 und 2.34) in x- und y-Richtung liefern, berechnet werden können.

2.5. Kapitelzusammenfassung

In diesem Kapitel wurden einige grundlegende mathematische und geometrische Zusammenhänge bei der Abbildung einer dreidimensionalen Szene in der Bildebene einer Kamera dargestellt, die für die Betrachtungen und das bessere Verständnis in den folgenden Kapiteln essentiell sind. Im Speziellen wurden dabei zuerst die Eigenschaften der projektiven Geometrie und die Projektion von Punkten in $\mathbb{R}^3$ auf eine Bildebene einer Kamera mithilfe der Zentralprojektion erläutert. Ferner wurde die Epipolargeometrie eingeführt, welche den Zusammenhang zwischen zwei Kameraansichten beschreibt. Abschließend wurde die Homographie als 2D-Transformation bzw. als perspektivische Transformation von Punkten zwischen zwei Ebenen dargestellt.

Das Kapitel hat nicht den Anspruch der Vollständigkeit. Vielmehr soll es den Einstieg in den Umgang der dreidimensionalen Bildverarbeitung geben. In den folgenden Kapiteln wird auf diesen Grundlagen lediglich aufgebaut und tiefgründigere Betrachtungen angestellt. Dennoch sei an dieser Stelle auf die einschlägige Literatur wie Hartley und Zisserman [68] und Faugeras und Luong [39] verwiesen, in der die aufgeführten Grundlagen vollständig dargestellt werden.

3. Robuste Schätzverfahren

In vielen modellbasierten Anwendungszenarien müssen die Modellparameter durch mehr oder weniger exakte Messungen ermittelt oder geschätzt werden. Die Genauigkeit dieser Schätzungen hängt hauptsächlich von der Exaktheit der Messungen ab. Ungenaue Messergebnisse bzw. vereinzelt falsche Messergebnisse können signifikante Auswirkungen auf die zu schätzenden Modellparameter haben.

Beispiel: Ein einfaches Beispiel ist in Abbildung 3.1 gegeben, bei dem ein linearer Zusammenhang zwischen zwei Messgrößen vorliegt. Die gestrichelte Gerade gibt jeweils den exakten - im Vorfeld nicht bekannten - Zusammenhang der Größen an. Die Messung sei nun fehlerbehaftet. Für die Bestimmung der Modellparameter - und damit der Geradengleichung - sind mindestens zwei Punkte notwendig. Im linken Koordinatensystem der Abbildung wurden zwei mögliche Geraden (blau und grau) bestimmt, indem jeweils willkürlich zwei Messpunkte ausgewählt wurden. Die Ergebnisse sind dabei äußerst verschieden. Bei der Wahl der beiden grauen Messpunkte liegt die geschätzte Gerade nahe an der wahren Geraden (gestrichelte Linie). Für die blauen Messpunkte hingegen weicht das Ergebnis sehr stark von dem wahren Verlauf ab. Die Ursache liegt in dem sehr großen Messfehler der Messung für den oberen blauen Punkt. Das Beispiel zeigt, dass das Ergebnis sehr stark von der Wahl der Messpunkte abhängt und somit nicht robust ist.

Zwei weitere Möglichkeiten, die Parameter der Geradengleichung zu bestimmen, sind im rechten Koordinatensystem der Abbildung 3.1 dargestellt. Werden alle Messungen (Punkte) herangezogen und die Parameter mittels des kleinsten quadratischen Fehlers (engl. *least squared error*, LSE) berechnet, so ergibt sich der durch die rote Linie dargestellte Verlauf, welcher wiederum - verursacht durch den großen Messfehler (roter Punkt) - stark von dem wahren Verlauf (gestrichelte Linie) abweicht. Wird die Messung mit dem starken Messfehler hingegen eliminiert und anschließend die Ausgleichsgerade über die Methode des kleinsten quadratischen Fehlers bestimmt, so liegt die Gerade (durchgezogene schwarze Linie) sehr nahe an der wahren Geraden.

Das vorrangige Ziel robuster Schätzverfahren liegt somit in der Eliminierung von sehr ungenauen und falschen Messergebnissen, die im Weiteren allgemein als Ausreißer (engl.

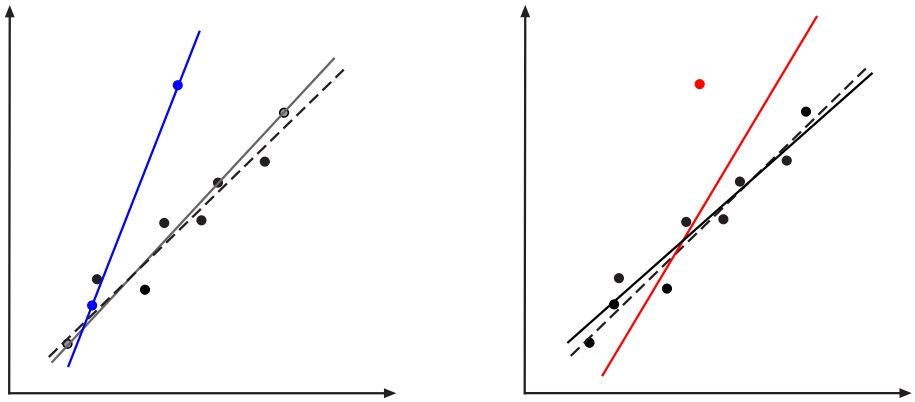


Abbildung 3.1.: Bestimmung einer Ausgleichsgraden

outlier) bezeichnen werden. Ferner wird bei der Vorstellung und Untersuchung robuster Schätzverfahren speziell auf die Bestimmung der Epipolargeometrie und damit auf die Schätzung der Fundamentalmatrix eingegangen, da diese essentiell für die weiteren Betrachtungen in dieser Arbeit ist. In Kapitel 4 werden noch einmal genauere Implementierungsdetails bezüglich eines verwendeten Algorithmus für die robuste Schätzung der Fundamentalmatrix und der 3D-Struktur dargelegt.

In den folgenden Abschnitten werden zunächst die grundlegenden Ansätze zur Schätzung der Fundamentalmatrix vorgestellt, die sich in *lineare Schätzverfahren* und *nichtlineare Schätzverfahren* unterteilen lassen. Robuste Ansätze, welche die Ausreißer eliminieren sollen und somit als Vorstufe der linearen und nichtlinearen Schätzverfahren angesehen werden können, werden in Abschnitt 3.3 beschrieben. In Abschnitt 3.3.2 werden zwei Fehlermaße definiert, die für den Vergleich und die Auswertung der Schätzverfahren in Abschnitt 3.4 Verwendung finden.

3.1. Lineare Schätzverfahren

In Abschnitt 2.3 wurden bereits die grundlegenden Eigenschaften der Fundamentalmatrix, welche die algebraische Darstellung der Epipolargeometrie ist, beschrieben. Die Epipolarbedingung ist definiert mit

$$\mathbf{m}'F\mathbf{m} = 0 \tag{3.1}$$

für korrespondierende Punkte $\{\mathbf{m}_i \leftrightarrow \mathbf{m}'_i\}$ in zwei Bildern.

3.1.1. 8-Punkt-Algorithmus

Für eine genügend große Anzahl an Punktkorrespondenzen - in der Regel werden acht Punktkorrespondenzen benötigt, mindestens aber sieben (siehe Abschnitt 2.3.3) - lässt

sich die Fundamentalmatrix bis auf einen Skalierungsfaktor berechnen. Der 8-Punkt-Algorithmus wurde erstmals von Longuet-Higgins [112] vorgestellt und wird allgemein auch als DLT-Algorithmus (engl. *direct linear transformation*) bezeichnet [68]. Mit

$$F = \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \tag{3.2}$$

und $\mathbf{m} = (x, y, 1)^T$ bzw. $\mathbf{m}' = (x', y', 1)^T$ lässt sich Gl. 3.1 umformulieren zu

$$x'xf_{11} + x'yf_{12} + x'f_{13} + y'xf_{21} + y'yf_{22} + y'f_{23} + xf_{31} + yf_{32} + f_{33} = 0, \tag{3.3}$$

bzw. in Vektorschreibweise

$$\left(x'x, x'y, x', y'x, y'y, y', x, y, 1\right) \mathbf{f} = 0. \tag{3.4}$$

Der Vektor $\mathbf{f} = (f_{11}, f_{12}, f_{13}, f_{21}, f_{22}, f_{23}, f_{31}, f_{32}, f_{33})^T$, welcher die Einträge der zu ermittelnden Fundamentalmatrix enthält, lässt sich über ein lineares Gleichungssystem berechnen. Mit n Punktkorrespondenzen in beiden Bildern kann Gl. 3.1 bzw. Gl. 3.4 in der Form

$$A\mathbf{f} = \begin{bmatrix} x'_1x_1 & x'_1y_1 & x'_1 & y'_1x_1 & y'_1y_1 & y'_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_nx_n & x'_ny_n & x'_n & y'_nx_n & y'_ny_n & y'_n & x_n & y_n & 1 \end{bmatrix} \mathbf{f} = 0 \tag{3.5}$$

aufgestellt werden.

Für eine eindeutige Lösung des Gleichungssystems müsste der Rang der Matrix A genau 8 sein. Für verrauschte Koordinaten der Punktkorrespondenzen, von denen im Allgemeinen ausgegangen werden muss, ist für $n > 8$ der Rang der Matrix A entsprechend ihrer Spaltenanzahl gleich 9.

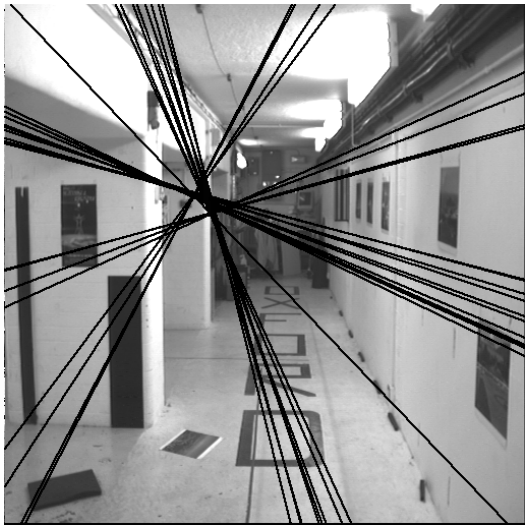
Definition 3.1.1 *Der rechte Nullraum der Matrix A - ein Unterraum gekennzeichnet mit null(A) - ist die Menge aller Lösungen für* $\mathbf{f}$ *des Gleichungssystems*

$$A\mathbf{f} = 0.$$

Die Lösung des Gleichungssystems wird mittels Singulärwertzerlegung (engl. *singular value decomposition*, SVD) ermittelt:

$$A = U\Sigma V^T. \tag{3.6}$$

Der kleinste quadratischer Fehler entspricht dem Vektor, dem der kleinste Singulärwert der Matrix A zugeordnet ist, d.h. der letzten Spalte von V. Algebraisch wird somit $\mathbf{f}$ durch Minimierung von $\|A\mathbf{f}\|$, mit $\|\cdot\|$ als l^2-Norm, unter der Bedingung, dass $\|\mathbf{f}\| = 1$ ist, ermittelt.



(a) F ist nicht singulär: Die Epipolarlinien schneiden sich nicht im Epipol. (Quelle: [68])

(b) F ist singulär: Die Epipolarlinien schneiden sich im Epipol. (Quelle: [68])

Abbildung 3.2.: Betrachtung der Singularitäts-Bedingung der Fundamentalmatrix anhand der Epipolarlinien

Singularität. Implizit ist mit der vorangehenden Berechnung noch nicht sichergestellt, dass die gefundene Fundamentalmatrix F den Rang 2 hat und somit singulär ist, welches aber zu den wesentlichen Eigenschaften dieser Matrix zählt. Für den Fall, dass die Fundamentalmatrix nicht singulär ist, schneiden sich die Epipolarlinien nicht in einem Punkt, dem Epipol, wie in Abbildung 3.2(a) dargestellt ist.

Um die Singularität zu erzwingen, wird nun über die Frobenius-Norm nach einer singulären Lösung für F gesucht, die am dichtesten an der zuvor berechneten Fundamentalmatrix liegt. Die Matrix F wird durch die Matrix $\hat{F}$ ersetzt, welche die Frobenius-Norm $\left\|F - \hat{F}\right\|$ in Hinblick auf $\mathbf{det}\hat{F} = 0$ minimiert. Wird für $F = U\Sigma V^T$ eine Singulärwertzerlegung durchgeführt, wobei $\Sigma = diag(\sigma_1, \sigma_2, \sigma_3)$ eine Diagonalmatrix ist mit $\sigma_1 > \sigma_2 > \sigma_3$, dann minimiert $\hat{F} = Udiag(\sigma_1, \sigma_2, 0)V^T$ die Frobenius-Norm. $\hat{F}$ hat den Rang 2 und ist damit singulär. In Abbildung 3.2(b) ist das Ergebnis für die singuläre Matrix $\hat{F}$ dargestellt. Der 8-Punkt-Algorithmus ist in Algorithmus 1 noch einmal zusammengefasst.

3.1.2. Normalisierter 8-Punkt-Algorithmus

Eine sinnvolle Erweiterung des 8-Punkt-Algorithmus wird über eine Normierung der Eingangsdaten erzielt. Die Lösung von Gl. 3.5 über den kleinsten quadratischen Fehler kann

Ziel: Für eine gegebene Anzahl $n \geq 8$ von Punktkorrespondenzen in zwei Bildern $\{\mathbf{m}_i \leftrightarrow \mathbf{m}'\}$ soll die Fundamentalmatrix F bestimmt werden, so dass die Epipolarbedingung $\mathbf{m}'^T F\mathbf{m} = 0$ erfüllt wird.

- Bestimme für alle Punktkorrespondenzen $\mathbf{m}_i \leftrightarrow \mathbf{m}'$ die Matrix A mittels Gleichung 3.5.
- Führe eine SVD von A durch. Der Spaltenvektor, der dem kleinsten Singulärwert zugeordnet ist, entspricht der Lösung von $\mathbf{f}$. Ist $A = U\Sigma V^T$ mit Σ als Diagonalmatrix mit positiven Singulärwerten, welche in absteigender Reihenfolge in der Diagonalen eingetragen sind, dann ist die Lösung für $\mathbf{f}$ die letzte Spalte von V.
- Setze die Fundamentalmatrix F mittels Gleichung 3.2 aus $\mathbf{f}$ zusammen.
- Erzwinge die Singularität der Matrix F (die Fundamentalmatrix muss den Rang 2 haben) durch Minimierung der Frobenius-Norm $\left\|F - \hat{F}\right\|$ mit $\mathbf{det}\hat{F} = 0$.
  1. Führe dazu erneut eine SVD durch, so dass $F = U\Sigma V^T$ ist.
  2. Setze den letzten Eintrag der Diagonalmatrix auf Null: $\Sigma = diag(\sigma_1, \sigma_2, 0)$.
  3. Die Lösung $\hat{F} = Udiag(\sigma_1, \sigma_2, 0)V^T$ ist die gesuchte singuläre Lösung für F.

Ergebnis: Die Matrix $\hat{F}$, für die der Vektor $\mathbf{f}$ das Gleichungssystem $\|A\mathbf{f}\|$ unter der Bedingung $\|\mathbf{f}\| = 1$ minimiert.

Algorithmus 1: DLT-Algorithmus für die Bestimmung von F

bei verrauschten Datensätzen zu ungenauen Ergebnissen führen, da die Matrix A schlecht konditioniert ist, d.h. die Spalten von A weisen unterschiedliche Größenordnungen auf:

$$A = \begin{bmatrix} x'_1x_1 & x'_1y_1 & x'_1 & y'_1x_1 & y'_1y_1 & y'_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_nx_n & x'_ny_n & x'_n & y'_nx_n & y'_ny_n & y'_n & x_n & y_n & 1 \\ \approx 10^4 & \approx 10^4 & \approx 10^2 & \approx 10^4 & \approx 10^4 & \approx 10^2 & \approx 10^2 & \approx 10^2 & 1 \end{bmatrix}.$$

Die Normalisierung erfolgt über eine Verschiebung und Skalierung der Bildkoordinaten, so dass der Schwerpunkt der Punktkorrespondenzen im Ursprung des Bildes und deren mittlerer Abstand vom Schwerpunkt $\sqrt{2}$ entspricht. Der 8-Punkt-Algorithmus mit zusätzlicher Normierung der Daten ist in Algorithmus 2 zusammengefasst.

Ein wichtiger Hinweis für den normalisierten 8-Punkt-Algorithmus ist, dass die Erzwingung der Singularität vor der abschließenden Denormalisierung stattfinden sollte [66].

3.2. Nichtlineare Schätzverfahren

Im Gegensatz zu den linearen Methoden zur Bestimmung der Fundamentalmatrix, welche ausschließlich auf algebraischen Zusammenhängen basieren, werden bei nichtlinearen Ansätzen vor allem die geometrischen Relationen berücksichtigt, um eine optimale

Ziel: Für eine gegebene Anzahl $n \geq 8$ von Punktkorrespondenzen in zwei Bildern $\{\mathbf{m}_i \leftrightarrow \mathbf{m}'_i\}$ soll eine Fundamentalmatrix gefunden werden, so dass $\mathbf{m}'^T F \mathbf{m} = 0$ erfüllt ist.

- **Normalisierung**: Transformiere die Bildkoordinaten nach $\tilde{\mathbf{m}}_i = T\mathbf{m}_i$ und $\tilde{\mathbf{m}}'_i = T'\mathbf{m}'_i$, wobei T und T' normalisierende Transformationen mit einer Verschiebung und Skalierung sind, so dass der Schwerpunkt aller Punkte im Koordinatenursprung $(0,0)^T$ liegt und der Mittelwert der Entfernung aller Punkte zum Ursprung $\sqrt{2}$ beträgt:

$$T = \begin{bmatrix} \frac{\sqrt{2}}{\sigma^2} & 0 & -\frac{\sqrt{2}\mu_x}{\sigma^2} \\ 0 & \frac{\sqrt{2}}{\sigma^2} & -\frac{\sqrt{2}\mu_y}{\sigma^2} \\ 0 & 0 & 1 \end{bmatrix},$$

  mit (μ_x, μ_y) dem Mittelwert der Punkte und σ^2 der Varianz des Abstandes der Punkte zum Schwerpunkt.

- Finde die Fundamentalmatrix $\tilde{F}$, die durch die Korrespondenzen $\{\tilde{\mathbf{m}}_i \leftrightarrow \tilde{\mathbf{m}}'_i\}$ definiert wird.
  1. **Lineare Lösung**: Finde eine Lösung für $\tilde{F}$ mittels des DLT-Algorithmus 1.
  2. **Erzwingen der Singularität**: Ersetze $\tilde{F}$ durch $\hat{F}$ durch Minimierung der Frobenius-Norm $\|\tilde{F} - \hat{F}\|$ mit $\mathbf{det}\hat{F} = 0$ (siehe Algorithmus 1).

- **Denormalisierung**: Setze $F = T'^T \hat{F} T$. Die Matrix F ist die Fundamentalmatrix, die die Epipolarbedingung $\mathbf{m}'^T F \mathbf{m} = 0$ erfüllt.

Ergebnis: Die Matrix F, für die der Vektor $\mathbf{f}$ das Gleichungssystem $\|A\mathbf{f}\|$ unter der Bedingung $\|\mathbf{f}\| = 1$ minimiert.

Algorithmus 2: Der normalisierte 8-Punkt-Algorithmus

Lösung für F zu finden. Da die Epipolarbedingung in Gl. 3.1 nichts weiter aussagt, als dass Punkte in einem Bild auf der korrespondierenden Epipolarlinie liegen müssen, wäre beispielsweise der Abstand der Punkte zu den Epipolarlinien ein geometrisches Fehlermaß, welches es zu minimieren gilt. Die verschiedenen Einträge der Fundamentalmatrix sind unterschiedlich bzw. nichtlinear abhängig von dem zugrunde liegenden Datensatz (Punktkorrespondenzen), weshalb lineare Methoden im eigentlichen Sinne nicht optimal sind. Ein typisches Energiefunktional E, welches in Abhängigkeit von den Parametern der Fundamentalmatrix minimiert werden soll, ist z. B. der *symmetrische Epipolarabstand* [202]:

$$\begin{aligned} E(F) &= \sum_{\forall i} d\left(\mathbf{m}'_i, F\mathbf{m}_i\right)^2 + d\left(\mathbf{m}_i, F^T\mathbf{m}'_i\right)^2 \\ &= \sum_{\forall i} \left(\mathbf{m}'^T_i F \mathbf{m}_i\right)^2 \left(\frac{1}{(F\mathbf{m}_i)_1^2 + (F\mathbf{m}_i)_2^2} + \frac{1}{\left(F^T\mathbf{m}'_i\right)_1^2 + \left(F^T\mathbf{m}'_i\right)_2^2}\right). \end{aligned} \tag{3.7}$$

$(\cdot)^2_{j|j=1,2}$ kennzeichnet das j-te Element des resultierenden Vektors. Mit $\mathbf{p} = (F_{11}, F_{12}, \ldots, F_{33})^T$ und $g_i(\mathbf{p}) = \sqrt{d\left(\mathbf{m}_i', F\mathbf{m}_i\right)^2 + d\left(\mathbf{m}_i, F^T\mathbf{m}_i'\right)^2}$ lässt sich das Energiefunktional vereinfacht umformulieren zu

$$E(\mathbf{p}) = \sum_{\forall i} \left(g_i(\mathbf{p})\right)^2. \tag{3.8}$$

Zur Lösung des nichtlinearen Minimierungsproblems existieren zahlreiche numerische Ansätze, die im Folgenden kurz beschrieben werden.

3.2.1. Gradientenverfahren

Das Gradientenverfahren, auch Gradientenabstiegsverfahren genannt, geht von einem Initialisierungswert des Parametersatzes aus und nähert sich schrittweise entlang dem negativen Gradienten hin zum Minimum, d.h. bis keine numerische Verbesserung mehr auftritt. Die Minimierung des Energiefunktionals erfolgt über den Gradienten

$$\begin{aligned} 0 &= \nabla_{\mathbf{p}} E(\mathbf{p}) \\ &= \frac{\partial E(\mathbf{p})}{\partial \mathbf{p}} \\ &= \left(\frac{\partial E}{\partial p_1}, \ldots, \frac{\partial E}{\partial p_n}\right)^T \\ &= 2 \sum_{\forall i} g_i(\mathbf{p}) \frac{\partial g_i(\mathbf{p})}{\partial \mathbf{p}}, \end{aligned} \tag{3.9}$$

wobei $\nabla_{\mathbf{p}}$ der Nabla-Operator mit den partiellen Ableitungen nach p_j darstellt. Ausgehend von einem initialen Parameter-Vektor $\mathbf{p}_0$ und dem Gradienten in Gl. 3.9 nähert sich das Verfahren schrittweise zum lokalen Minimum mit

$$\mathbf{p}_{k+1} = \mathbf{p}_k - \delta_k \nabla_{\mathbf{p}} E\left(\mathbf{p}_k\right) \quad mit\ k \geq 0, \tag{3.10}$$

so dass $E(\mathbf{p}_k) > E(\mathbf{p}_{k+1})$. δ_k gibt die Schrittweite des Verfahrens an, welche sich über den Betrag des Gradienten bestimmen lässt. Das Gradientenverfahren konvergiert zwar, ist allerdings sehr langsam, gerade wenn es sich dem Optimum nähert und dort der Betrag des Gradienten und damit die Schrittweite sehr klein wird.

3.2.2. Gauß-Newton-Verfahren

Das *Gauß-Newton*-Verfahren, welches eine Modifikation des Newton-Verfahrens darstellt, ist ebenfalls ein iteratives Optimierungsverfahren zur Lösung nichtlinearer Minimierungsprobleme. Im Gegensatz zum Newton-Verfahren, welches hier nicht näher ausgeführt wird, benötigt es allerdings nicht die zweite Ableitung des Energiefunktionals. Diese wird

über eine Taylorreihe erster Ordnung linear approximiert, weshalb die Aufstellung der Hessematrix vermieden wird.

Ausgehend von einer initialen Schätzung des Parametersatzes werden die Parameter schrittweise nachgeregelt, so dass Gl. 3.9 einem lokalen Minimum zustrebt mit

$$\begin{aligned} \mathbf{p}_{k+1} &= \mathbf{p}_k - \left[J_g(\mathbf{p}_k)^T J_g(\mathbf{p}_k) \right]^{-1} J_g(\mathbf{p}_k) g(\mathbf{p}_k) \\ &= \mathbf{p}_k + \Delta \mathbf{p}_k \quad mit\ k \geq 0. \end{aligned} \tag{3.11}$$

$J_g(\mathbf{p}_k)$ ist die Jakobimatrix

$$J_g(\mathbf{p}_k) = \begin{bmatrix} \frac{\partial (g_1(\mathbf{p}))^2}{\partial p_1} & \frac{\partial (g_1(\mathbf{p}))^2}{\partial p_2} & \cdots & \frac{\partial (g_1(\mathbf{p}))^2}{\partial p_n} \\ \frac{\partial (g_2(\mathbf{p}))^2}{\partial p_1} & \frac{\partial (g_2(\mathbf{p}))^2}{\partial p_2} & \cdots & \frac{\partial (g_2(\mathbf{p}))^2}{\partial p_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial (g_m(\mathbf{p}))^2}{\partial p_1} & \frac{\partial (g_m(\mathbf{p}))^2}{\partial p_2} & \cdots & \frac{\partial (g_m(\mathbf{p}))^2}{\partial p_n} \end{bmatrix} \tag{3.12}$$

und $\Delta \mathbf{p}_k$ der Aktualisierungsvektor

$$\Delta \mathbf{p}_k = - \left[J_g(\mathbf{p}_k)^T J_g(\mathbf{p}_k) \right]^{-1} J_g(\mathbf{p}_k) g(\mathbf{p}_k), \tag{3.13}$$

welcher über die Normalengleichungen [57]

$$J_g(\mathbf{p}_k)^T J_g(\mathbf{p}_k) \Delta \mathbf{p}_k = -J_g(\mathbf{p}_k) g(\mathbf{p}_k) \tag{3.14}$$

durch *QR-Zerlegung* ohne Berechnung der Inversen ermittelt werden kann.

Nachteilig bei der Gauß-Newton-Methode ist, dass das Verfahren nicht zwingend konvergiert, wenn die initiale Lösung zu weit vom Optimum entfernt liegt [162].

3.2.3. Levenberg-Marquardt Verfahren

Ein weit verbreitetes und im Gegensatz zum Gauß-Newton-Algorithmus robusteres Optimierungsverfahren ist der *Levenberg-Marquardt*-Algorithmus (LM) [127], welcher als eine Kombination aus dem Gradientenverfahren und dem Gauß-Newton-Verfahren betrachtet werden kann. Der LM-Algorithmus verhält sich bei schlechter Initialisierung wie das Gradientenverfahren. Liegt die aktuelle Lösung dicht am Optimum, bei dem das Gradientenverfahren sich weiter verlangsamt, wechselt der Algorithmus zur Gauß-Newton-Methode.

Das grundlegende Merkmal des LM-Algorithmus liegt in der Variation von Gl. 3.11 bzw. Gl. 3.14 durch Erweiterung um einen Dämpfungsterm $\lambda > 0$:

$$\mathbf{p}_{k+1} = \mathbf{p}_k - \left[J_g(\mathbf{p}_k)^T J_g(\mathbf{p}_k) + \lambda I \right]^{-1} J_g(\mathbf{p}_k) g(\mathbf{p}_k) \tag{3.15}$$

bzw.

$$\left[J_g(\mathbf{p}_k)^T J_g(\mathbf{p}_k) + \lambda I \right] \Delta \mathbf{p}_k = -J_g(\mathbf{p}_k) g(\mathbf{p}_k). \tag{3.16}$$

Die außerhalb der Diagonalen liegenden Elemente der Matrix $\left[J_g(\mathbf{p}_k)^T J_g(\mathbf{p}_k)\right]$ bleiben unverändert. Nur die diagonalen Elemente $\left[J_g(\mathbf{p}_k)^T J_g(\mathbf{p}_k)\right]_{ii}$ werden um den Dämpfungsterm λ ergänzt. Verringert sich das Energiefunktional $E(\mathbf{p})$ in einem Iterationsschritt mit dem neuen Parametervektor $\mathbf{p}_{k+1}$ ($E(\mathbf{p}_{k+1}) < E(\mathbf{p}_k)$), dann wird dieser übernommen und λ reduziert. Auf der anderen Seite wird der Dämpfungsterm vergrößert, wenn das Energiefunktional und damit der Fehler zunimmt. Der aktualisierte Parameter-Vektor wird dann in diesem Iterationsschritt nicht übernommen und mit dem neuen λ noch einmal berechnet. Durch die Erhöhung des Dämpfungsterms verstärkt sich die Wirkung der Elemente auf der Hauptdiagonalen $\left[J_g(\mathbf{p}_k)^T J_g(\mathbf{p}_k)\right]_{ii}$. Unter Vernachlässigung der Nebenelemente tendiert der LM-Algorithmus dann zum Gradientenverfahren. Bei einer fortschreitenden Reduzierung des Dämpfungsterms verschwindet der Term λI und Gl. 3.15 und Gl. 3.11 sind identisch, d.h. der LM-Algorithmus tendiert zum Gauß-Newton-Verfahren. Die Berechnungsschritte des LM-Algorithmus sind im Anhang A.2 zusammengefasst.

3.3. Robuste Schätzverfahren

Die in den vorangegangenen Abschnitten beschriebenen linearen und nichtlinearen Schätzverfahren haben den Nachteil, dass sie bei schlechten Messergebnissen, verursacht durch nicht normalverteilte Rauscheinflüsse oder einfach durch fehlerhafte Messungen, ungenaue Ergebnisse liefern bzw., im Falle der nichtlinearen Schätzverfahren, durch eine schlechte Initialisierung anstelle des globalen Minimums nur das lokale Minimum der verwendeten Kostenfunktion finden. Robuste Verfahren hingegen werden vor allem dazu verwendet, Ausreißer zu eliminieren und einen initialen Schätzwert zu ermitteln, der möglichst in der Nähe des globalen Minimums liegt. Für Punktkorrespondenzen, die für die Schätzung der Fundamentalmatrix verwendet werden, unterscheidet Zhang in [202] zwei Arten von Ausreißern: schlechte Punktkorrespondenzen, deren Koordinaten stark verrauscht sind, und falsche Punktkorrespondenzen, die nicht den selben Raumpunkt abbilden.

In der Literatur sind viele robuste Ansätze zu finden [23, 39, 47, 48, 72, 132, 153, 184, 186, 202], die unterschiedliche Strategien verfolgen und auf verschiedenen Annahmen basieren. Der M-Schätzer (engl. *M-estimator*) [72], welcher auf einer unterschiedlichen Gewichtung der einzelnen Messergebnisse beruht, stellt davon das wohl bekannteste robuste Schätzverfahren dar. Allerdings wird der M-Schätzer aufgrund seiner geringen Robustheit bei falschen Punktkorrespondenzen zur Bestimmung der Fundamentalmatrix nicht oder kaum noch verwendet.

Ein weiteres - ebenfalls sehr bekanntes - Schätzverfahren ist die Methode des *kleinsten Medianwertes des quadratischen Fehlers* (engl. *least median of squares*, LMedS). Dieses

Schätzverfahren ist, obwohl ebenfalls kaum noch zur Schätzung der Fundamentalmatrix verwendet, im Vergleich zum M-Schätzer wesentlich robuster bezüglich des Auftretens falscher Punktkorrespondenzen, da es diese größtenteils bei der Berechnung eliminiert. Die LMedS Methode zählt u.a. zu den robusten Schätzverfahren, welche auch als *Monte-Carlo*-Methoden[1] bekannt sind. Monte-Carlo-Methoden, zu denen auch der RANSAC- (engl. ***RAN**dom **SA**mpling **C**onsensus*) [48], MLESAC- (engl. ***M**aximum **L**ikelihood **E**stimation **SA**mple **C**onsensus*) [186], MAPSAC- (engl. ***M**aximum **A P**ostiori **SA**mple **C**onsensus*) [184], NAPSAC- (engl. ***N A**djacent **P**oints **SA**mple **C**onsensus*) [132] und PROSAC- (engl. ***P**rogressive **SA**mple **C**onsensus*) [23] Algorithmus zählt, sind die wohl am häufigsten verwendeten Verfahren zur Schätzung der Epipolargeometrie, von denen einige im folgenden Abschnitt näher beschrieben werden.

3.3.1. Monte-Carlo-Methoden

Der grundlegende Gedanke von Monte-Carlo-Methoden ist, dass aus der Menge der zur Verfügung stehenden Daten zufällig ein Satz an Daten ausgewählt wird, der mindestens für die Berechnung der zu schätzenden Parameter notwendig ist. Ausgehend von einem Datensatz n korrespondierender Punkte in zwei Ansichten $\mathcal{S} = \{(\mathbf{m}_i \leftrightarrow \mathbf{m}'_i) \,|i = 1, 2, ..., n\}$ wird demzufolge zufällig eine Untermenge $\mathcal{S}_j \subseteq \mathcal{S} :\Leftrightarrow \forall \mathbf{m} : \mathbf{m} \in \mathcal{S}_j \Rightarrow \mathbf{m} \in \mathcal{S}$ mit s Korrespondenzen ausgewählt, um die Parameter der Fundamentalmatrix - z. B. über den normalisierten 8-Punkt-Algorithmus - zu berechnen. Die zufällige Auswahl einer Untermenge wird für eine Anzahl N wiederholt, so dass mit einer Wahrscheinlichkeit p mindestens eine der zufällig ausgewählten Untermengen frei von Ausreißern ist. Die Anzahl N an Wiederholungen berechnet sich mit

$$N = \frac{\log(1-p)}{\log(1-(1-\epsilon)^s)}, \tag{3.17}$$

wobei ϵ die Wahrscheinlichkeit angibt, dass eine zufällig ausgewählte Punktkorrespondenz ein Ausreißer ist. Wird der normalisierte 8-Punkt-Algorithmus zur Berechnung von F verwendet, ist $s = 8$ zu setzen, da - wie in Abschnitt 3.1.1 beschrieben - 8 Punktkorrespondenzen ausreichend sind, um eine eindeutige Lösung für das Gleichungssystems in Gl. 3.5 zu finden. Für p wird in der Regel ein Wert von $0,99$ verwendet. Die Anzahl der Ausreißer und damit die Wahrscheinlichkeit ϵ, dass eine zufällig ausgewählte Punktkorrespondenz ein Ausreißer ist, sind im Vorfeld nicht bekannt. In der Initialisierungsphase wird für ϵ ein Wert eingesetzt, der oberhalb des anzunehmenden prozentualen Anteils an Ausreißern liegt.

[1] Monte-Carlo-Methoden wurden ursprünglich unter dem Namen „statistisches Sampling“ bzw. „statistische Stichprobenentnahme“ geführt. Die Art des Zufallsprozesses ist analog zu den Prozessen, wie sie z. B. in Spielcasinos auftreten.

Beispiel: Seien $p = 0,99$ und $s = 8$ gegeben, so ergibt sich für N in Abhängigkeit von ϵ: $N = 9|_{\epsilon=0,1}$, $N = 26|_{\epsilon=0,2}$, $N = 78|_{\epsilon=0,3}$, $N = 272|_{\epsilon=0,4}$, $N = 1177|_{\epsilon=0,5}$ und $N = 7025|_{\epsilon=0,6}$.

In jedem Iterationsschritt wird für den gesamten Datensatz der Fehler (z. B. der symmetrische Epipolarabstand aus Gl. 3.7) ermittelt und diejenigen Punktkorrespondenzen als Ausreißer klassifiziert, die außerhalb eines definierten Toleranzbereichs liegen, d.h. dessen Abstand zu den korrespondierenden Epipolarlinien einen Schwellwert überschreiten. Nach N Durchgängen wird diejenige Lösung der Parameter ausgewählt, die das beste Ergebnis in Hinblick auf ein verwendetes Kostenfunktional liefert.

In den folgenden Abschnitten werden mehrere Verfahren vorgestellt, die diesen Monte-Carlo-Ansatz verfolgen. Allerdings unterscheiden sich die Verfahren sowohl in der Bestimmung des Schwellwertes zur Ausreißer-Eliminierung als auch in der Wahl des verwendeten Kostenfunktionals. Als quadratisches Fehlermaß wird beispielhaft der in Gl. 3.7 angegebene symmetrische Epipolarabstand der einzelnen Punktkorrespondenzen

$$e_i^2 = d\left(\mathbf{m}_i', F\mathbf{m}_i\right)^2 + d\left(\mathbf{m}_i, F^T\mathbf{m}_i'\right)^2 \tag{3.18}$$

verwendet. In Abschnitt 3.3.2 werden zwei weitere Fehlermaße eingeführt, die in Abschnitt 3.4 zum Vergleich der robusten Schätzverfahren anstelle des symmetrischen Epipolarabstands eingesetzt wurden. An dieser Stelle sei darauf hingewiesen, dass die Monte-Carlo-Verfahren allgemein nur als Initialisierung angesehen werden sollen. Nach der Eliminierung von Ausreißern sind weitere Verfahren, wie in den Abschnitten 3.1 bzw. 3.2 beschrieben, zur Optimierung anzuwenden.

Least Median of Squares (LMedS)

Der LMedS-Algorithmus [39, 153] berechnet in jedem Durchlauf $j = \{1, 2, \ldots, N\}$ die Fundamentalmatrix F_j mit dem minimalen Satz an Punktkorrespondenzen und den Medianwert für den gesamten Datensatz mit

$$E_j = \underset{i=1,\ldots,n}{\text{median}}\, e_i^2. \tag{3.19}$$

Nach N Iterationen wird diejenige Fundamentalmatrix F_j ausgewählt, für die E_j den kleinsten von allen Medianwerten besitzt. Die Eliminierung von Ausreißern erfolgt über die Berechnung eines Schwellwertes $T = (2,5\hat{\sigma})^2$ [202] bzw. $T = 3,84\hat{\sigma}^2$ [68], wobei $\hat{\sigma}$ die geschätzte *robuste Standardabweichung* [202]

$$\hat{\sigma} = 1,4826\,[1 + 5/(n-s)]\,\underset{i=1,\ldots,n}{\text{median}}\,|e_i| \tag{3.20}$$

ist. Der Wert $1,4826$ ist dabei ein Koeffizient, der denselben Effekt erzielen soll wie bei der Lösung des kleinsten quadratischen Fehlers, wenn ausschließlich gaußverteiltes Rauschen auf den Daten liegt. Der Term $5/(n-s)$ dient zur Kompensation von sehr kleinen

Datensätzen, wobei n der Anzahl der Punktkorrespondenzen und s der Anzahl der zu schätzenden Parameter entspricht. An dieser Stelle wird für nähere Informationen über diese Werte auf [153] verwiesen.

Die Aufteilung des Datensatzes in Punktkorrespondenzen, dessen quadratischer Fehler kleiner als der Schwellwert ist, und Ausreißer erfolgt über eine Indizierung oder Wichtung

$$w_i = \begin{cases} 1 & wenn\ e_i^2 \leq T \\ 0 & sonst \end{cases} . \tag{3.21}$$

Da der Anteil an Ausreißern ϵ im Vorfeld, wie bereits erwähnt, nicht bekannt ist, kann optional in jedem Durchlauf j über Gl. 3.20 und Gl. 3.21 der Ausreißeranteil bestimmt und damit N neu berechnet werden. Die Anzahl an Durchläufen kann hierdurch reduziert werden, da in der Regel in der Initialisierungsphase ein zu hoher Wert für ϵ verwendet wird.

Random Sampling Consensus (RANSAC)

Die Funktionsweise des RANSAC-Algorithmus ähnelt sehr stark dem LMedS-Verfahren. Der wesentliche Unterschied liegt in der Auswahl der am besten geeigneten Fundamentalmatrix F_j. Wohingegen der LMedS-Algorithmus anhand des kleinsten Medianwertes eine Entscheidung trifft, wird bei RANSAC in jedem Durchlauf die Anzahl der Ausreißer über Gl. 3.21 ermittelt und diejenige Fundamentalmatrix ausgewählt, welche den kleinsten Ausreißeranteil bzw. die größte Anzahl an Punktkorrespondenzen, die innerhalb des Toleranzbereiches liegen, liefert.

Für einen großen Anteil an Ausreißern - vor allem falsche Zuordnungen von Punktepaaren - ist RANSAC gegenüber LMedS aufgrund der Robustheit vorzuziehen, da der Medianwert spätestens bei $\epsilon = 0,5$ nicht mehr geeignet ist. Eine detailliertere Beschreibung des RANSAC-Verfahrens mit einigen Erweiterungen ist in Kapitel 4 (3D-Rekonstruktion aus zwei Ansichten) Abschnitt 4.3 zu finden.

Maximum Likelihood Estimation Sample Consensus (MLESAC)

MLESAC [186] basiert auf der selben Sampling-Strategie, wie die zuvor beschriebenen LMedS und RANSAC. Der Unterschied liegt, wie auch zwischen LMedS und RANSAC, in der Auswahl für die beste Lösung F_j aus einem der Iterationsschritte. MLESAC verwendet eine Maximum-Likelihood-Schätzung, also eine Schätzung, die mit größter Wahrscheinlichkeit die Beste der N Schätzungen für die Parameter der Fundamentalmatrix ist, indem in jedem Durchlauf die Verteilung des Fehlers der Punktkorrespondenzen nachgebildet wird. Dabei wird die Annahme getroffen, dass das Rauschen auf den Koordinaten

richtiger Zuordnungen gaußverteilt ist und bei falschen Zuordnungen das Rauschen einer Gleichverteilung entspricht. Der Fehler lässt sich damit als ein Mischmodell einer Gauß- und Gleichverteilung über die Wahrscheinlichkeitsfunktion

$$p(e_i) = \gamma \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^D \exp\left(-\frac{e_i^2}{2\sigma^2} \right) + (1-\gamma)\,\frac{1}{\nu} \tag{3.22}$$

modellieren. γ ist die Wahrscheinlichkeit einer richtigen Zuordnung, σ ist die Standardabweichung und e_i^2 sind die symmetrischen Epipolarabstände aus Gl. 3.18. ν gibt die Größe des Bereichs an, innerhalb dessen Ausreißer entsprechend einer Gleichverteilung fallen und D ist die Dimension des Fehlers ($D = 4$, da $\{e_i^2\} \in \mathbb{R}^4$). Anstelle der Maximierung der Gl. 3.22 über alle Fehler $\{(e_i^2)|i = 1, 2, \ldots, n\}$ (Maximum-Likelihood) in jedem Durchlauf $j = \{1, 2, \ldots, N\}$ wird in [186] als Kostenfunktion die negative logarithmische Wahrscheinlichkeitsfunktion minimiert:

$$-L_j = -\sum_{\forall i} \log\left(\gamma \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^D \exp\left(-\frac{e_i^2}{2\sigma^2} \right) + (1-\gamma)\,\frac{1}{\nu} \right). \tag{3.23}$$

Da γ auch als Regelungsgröße des Mischmodells angesehen werden kann, erfolgt die Minimierung von Gl. 3.23 durch Schätzung von γ z. B. über den EM-Algorithmus (engl. *expectaion maximization*). Ausgehend von einer initialen Schätzung für $\hat{\gamma}_0$ wird iterativ $\hat{\gamma}_{k+1}$ ($k \geq 0$) über

$$\hat{\gamma}_{k+1} = \frac{1}{n} \sum_{i=1}^{n} z_i \tag{3.24}$$

mit

$$z_i = \frac{\hat{\gamma}_k \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^D \exp\left(-\frac{e_i^2}{2\sigma^2} \right)}{\hat{\gamma}_k \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^D \exp\left(-\frac{e_i^2}{2\sigma^2} \right) + (1-\hat{\gamma}_k)\,\frac{1}{\nu}} \tag{3.25}$$

solange neu berechnet, bis eine Konvergenz auftritt (in der Regel drei bis vier Iterationen, wenn als Startwert $\hat{\gamma}_0 = 0,5$ gewählt wird).

Abschließend wird diejenige Fundamentalmatrix F_j aus Durchlauf j ausgesucht, bei der die Kostenfunktion $-L_j$ den kleinsten Wert liefert.

Maximum A Posteriori Sample Consensus (MAPSAC)

Ähnlich dem MLESAC wird bei dem MAPSAC [184] die Verteilung des Fehlers der Punktkorrespondenzen über ein Mischmodell aus Gaußverteilung (richtige Zuordnungen) und Gleichverteilung (falsche Zuordnungen) entsprechend Gl. 3.22 modelliert. Die zu minimierende Kostenfunktion ist

$$E_j = \sum_{i=1}^{n} \rho\left(\frac{e_i^2}{\sigma^2} \right) \tag{3.26}$$

mit

$$\rho\left(\frac{e_i^2}{\sigma^2}\right) = \begin{cases} \frac{e_i^2}{\sigma^2} & wenn\ \frac{e_i^2}{\sigma^2} \leq T \\ T & sonst \end{cases} . \tag{3.27}$$

Der Schwellwert T mit

$$T = 2\log\left(\frac{\gamma}{1-\gamma}\right) + (D-d)\log\left(\frac{L^2}{2\pi\sigma^2}\right) \tag{3.28}$$

wird iterativ über eine Schätzung für γ entsprechend Gl. 3.24 und Gl. 3.25 bestimmt, so dass Gl. 3.26 minimal wird. d ist die Dimension der Mannigfaltigkeit ($d = 3$ für die Schätzung der Fundamentalmatrix) und L^2 gibt die Größe des Bildbereiches (Suchbereich, siehe Abschnitt 4.2) an, innerhalb dessen falsche Zuordnungen einer Gleichverteilung unterliegen. In [47] wird dieser Ansatz noch erweitert, indem gleichzeitig in jedem dieser Iterationsschritte die Standardabweichung mit

$$\hat{\sigma}_{k+1} = \sqrt{\frac{\sum_{i=1}^{n}\left(z_i \cdot e_i^2\right)}{D\sum_{i=1}^{n} z_i}} = \sqrt{\frac{\sum_{i=1}^{n}\left(z_i \cdot e_i^2\right)}{n\hat{\gamma}_{k+1}D}} \tag{3.29}$$

ermittelt wird. Als Startwerte werden in [47] für $\hat{\gamma}_0 = 0,5$ und $\hat{\sigma}_0 = \sqrt{\frac{\text{median}(e_i^2)}{D}}$ vorgeschlagen.

Abschließend wird wiederum diejenige Fundamentalmatrix F_j aus Durchlauf j ausgesucht, bei der die Kostenfunktion E_j den kleinsten Wert liefert.

3.3.2. Fehlermaße

Anstelle des symmetrischen Epipolarabstands werden in [68] zwei weitere Fehlermaße vorgestellt, welche in dieser Arbeit Verwendung finden: der Sampson-Fehler und der 2D-Reprojektionsfehler. Der Sampson-Fehler wird häufig mit dem symmetrischen Epipolarabstand gleichgestellt, da er ebenfalls auf den Abständen der Punkte zu den korrespondierenden Epipolarlinien basiert. Bei dem 2D-Reprojektionsfehler hingegen werden über die Fundamentalmatrix zuerst die Projektionsmatrizen der Kameras ermittelt und anschließend eine Triangulation durchgeführt, um die 3D-Raumkoordinaten $\mathbf{M}_i$ der Punktkorrespondenzen $\{\mathbf{m}_i \leftrightarrow \mathbf{m}_i'\}$ in den beiden Bildern zu bestimmen. Der 2D-Reprojektionsfehler ergibt sich somit aus den Abständen der Projektionen dieser Raumpunkte auf den Bildebenen $\{\hat{\mathbf{m}}_i,\ \hat{\mathbf{m}}_i'\}$ und den entsprechenden Punkten $\{\mathbf{m}_i,\ \mathbf{m}_i'\}$.

Sampson-Fehler

Der Sampson-Fehler stellt eine Approximation des geometrischen Fehlers durch Terme erster Ordnung dar [68]. Die Kostenfunktion, welche es zu minimieren gilt, lautet

$$E_{Samp}(F) = \sum_{\forall i} \frac{\left(\mathbf{m}_i'^T F \mathbf{m}_i\right)^2}{(F\mathbf{m}_i)_1^2 + (F\mathbf{m}_i)_2^2 + \left(F^T\mathbf{m}_i'\right)_1^2 + \left(F^T\mathbf{m}_i'\right)_2^2}. \tag{3.30}$$

2D-Reprojektionsfehler

Sowohl bei dem symmetrischen Epipolarabstand als auch bei dem Sampson-Fehler beschränkt sich die Minimierung der Kostenfunktion auf die Parameter der Fundamentalmatrix. Für den 2D-Reprojektionsfehler werden zusätzlich die Koordinaten der 3D-Raumpunkte benötigt, welche erst über Triangulationsverfahren ermittelt werden müssen. In Kapitel 4 werden einige verschiedene Arten der Triangulation näher beschrieben.

Der 2D-Reprojektionsfehler lässt sich mit

$$E_{rep}(F) = \sum_{\forall i} d(\mathbf{m}_i, \hat{\mathbf{m}}_i)^2 + d(\mathbf{m}'_i, \hat{\mathbf{m}}'_i)^2 \tag{3.31}$$

berechnen, wobei $d(\cdot,\cdot)$ in diesem Fall der Euklidische Abstand zwischen den Punkten $\mathbf{m}_i$ und $\hat{\mathbf{m}}_i = P\mathbf{M}_i$ im ersten Bild bzw. $\mathbf{m}'_i$ und $\hat{\mathbf{m}}'_i = P'\mathbf{M}_i$ im zweiten Bild ist.

3.4. Vergleich der robusten Schätzverfahren

In diesem Abschnitt werden die zuvor beschriebenen robusten Schätzverfahren (LMedS, RANSAC, MLESAC und MAPSAC) näher untersucht und verglichen. Der Vergleich ist dahingehend schwierig, da alle diese Verfahren zum einen Zufallsprozesse sind und zum anderen, wenn die wahre zugrunde liegende Epipolargeometrie nicht bekannt ist, Fehlerabstände nicht zwischen den wahren und den geschätzten Größen ermittelt werden können.

Der erste Punkt lässt sich durch zahlreiches Wiederholen der jeweiligen Zufallsprozesse lösen, so dass für das entsprechende Verfahren eine statistische Aussage getroffen werden kann. Für den zweiten Fall ist es sinnvoll, einen Datensatz zu simulieren, für den die Epipolargeometrie bekannt ist und der aus exakten Punktkorrespondenzen besteht. Die gauß- und gleichverteilten Rauscheinflüsse lassen sich dann nachträglich und kontrolliert auf diesen Datensatz anwenden.

3.4.1. Modellierung der Epipolargeometrie

Für den Vergleich der verschiedenen robusten Ansätze werden zwei Datensätze modelliert, indem jeweils eine Anzahl von 3D-Raumpunkten $\{\mathbf{M}_i | i = 1, 2, \ldots, n\}$ und die Projektionsmatrizen P und P' definiert werden. Sowohl die Fundamentalmatrix als auch die 2D-Punktkorrespondenzen in den beiden Bildebenen der Kameras lassen sich damit exakt erzeugen:

$$\begin{aligned} \mathbf{m}_i &= P\mathbf{M}_i \\ \mathbf{m}'_i &= P'\mathbf{M}_i \end{aligned}$$

bzw.

$$F = \left[P'\mathbf{C}\right]_{\times} P'P^{+} \tag{3.32}$$

mit $\mathbf{C}$ als Ursprung des Weltkoordinatensystems und P^{+} als Pseudoinverse der Matrix P.

3.4.2. Fehlermaße zum Vergleich der Ergebnisse

Zum Vergleich der Schätzergebnisse werden zwei Fehlermaße herangezogen. Das erste Vergleichsmaß ist der mittlere Abstand der Epipolarlinien zu den korrespondierenden Punkten:

$$e_1 = \frac{1}{2n}\sum_{i=1}^{n} d(\mathbf{m}_i', \hat{F}\mathbf{m}_i) + d(\mathbf{m}_i, \hat{F}^T\mathbf{m}_i'), \tag{3.33}$$

d.h. für alle robusten Verfahren werden die Fundamentalmatrizen $\hat{F}$ aus einem verrauschten und ggf. mit Ausreißern versehenen Datensatz mit den Punktkorrespondenzen $\{\tilde{\mathbf{m}}_i \leftrightarrow \tilde{\mathbf{m}}_i'\}$ geschätzt und anschließend die mittleren Abstände der Epipolarlinien zu dem fehlerfreien Datensatz $\{\mathbf{m}_i \leftrightarrow \mathbf{m}_i'\}$ berechnet.

Das zweite Fehlermaß ist der relative Abstand der Epipole. Dieses Vergleichsmaß wird von Luong und Faugeras [117] vorgeschlagen, wenn die wahre Fundamentalmatrix F bekannt ist. Da der Epipol implizit in der Fundamentalmatrix enthalten ist, stellt der Abstand der Epipole zwischen den geschätzten Matrizen $\hat{F}$ und der tatsächlichen Matrix F ein direktes Fehlermaß dar. Da die Auswirkung des Fehlers abhängig von der Lage der Epipole ist (d.h. je dichter der Epipol am Bildausschnitt bzw. am Kamerahauptpunkt liegt desto größer ist die Auswirkung des absoluten Fehlers), wird ein relativer Abstand definiert. Für die Koordinate $\hat{x}$ des geschätzten Epipols ist der relative Fehler bezüglich der Koordinate x des wahren Epipols

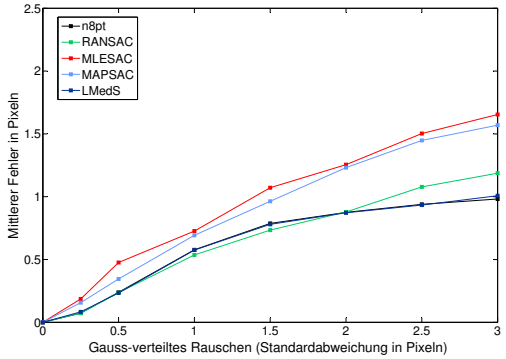
$$e_x = \min\left\{\frac{|x-\hat{x}|}{\min(|x|,|\hat{x}|)}, 1\right\} \tag{3.34}$$

im ersten Bild. Entsprechend Gl. 3.34 werden die relativen Fehler e_y für die y-Koordinate und $e_{x'}$ und $e_{y'}$ für das zweite Bild bestimmt. Der Gesamtfehler ergibt sich durch Summation der Einzelfehler zu

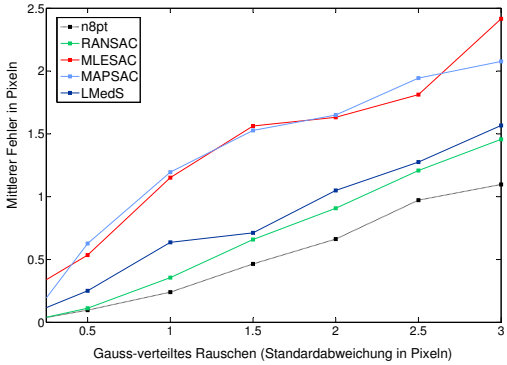
$$e_2 = e_x + e_y + e_{x'} + e_{y'}. \tag{3.35}$$

3.4.3. Ergebnisse und Auswertung

In diesem Abschnitt soll nun die Robustheit der Schätzverfahren (LMedS, RANSAC, MLESAC, MAPSAC) mit zwei zuvor erzeugten Testdatensätzen und Fehlermaßen untersucht werden. Die beiden Testdatensätze unterscheiden sich hauptsächlich in der Lage der Epipole. Für Datensatz 1 liegt der Epipol innerhalb des Bildes und für Datensatz 2 ist der Epipol weit außerhalb des Bildes. Die robusten Ansätze werden lediglich dazu verwendet, Ausreißer zu eliminieren. Im Anschluss an die Eliminierung wird der normalisierte 8-Punkt-Algorithmus auf die übrigen Punktkorrespondenzen angewendet.



(a) Verlauf der mittleren Abstände zwischen den berechneten Epipolarlinien und den korrekten Bildpunkten (Datensatz 1, Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400)



(b) Verlauf der mittleren Abstände zwischen den berechneten Epipolarlinien und den korrekten Bildpunkten (Datensatz 2, Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 500)

Abbildung 3.3.: Vergleich robuster Schätzer bei verrauschten Datensätzen: Steigerung der Standardabweichung des gaußverteilten Rauschens von 0 bis 3, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler verwendet.

Versuch 1: Im ersten Versuch wird auf die Koordinaten der beiden Testdatensätze gaußverteiltes Rauschen mit einer ansteigenden Standardabweichung $\sigma = \{0,25,\ 0,5,\ 1,0,\ 1,5,\ 2,0,\ 2,5,\ 3,0\}$ Pixel addiert und jeweils die Fundamentalmatrix mit den vier verschiedenen robusten Ansätzen ermittelt. Die zu minimierenden Kostenfunktionen basieren zum einen auf dem Sampson-Fehler und zum anderen auf dem 2D-Reprojektionsfehler. Für jeden der robusten Ansätze wird die Fundamentalmatrix 50 mal geschätzt, die Fehler e_1 und e_2 jeweils bestimmt und über die 50 Iterationen gemittelt. Da in diesem Versuch keine falschen Punktkorrespondenzen erzeugt werden, wird zusätzlich als Vergleich der normalisierte 8-Punkt-Algorithmus auf den gesamten Datensatz angewendet.

In Abbildung 3.3 sind die mittleren Abstände der Epipolarlinien zu den korrespondierenden Punkten in Abhängigkeit von der eingestellten Standardabweichung des gaußverteilten Rauschens für die beiden erzeugten Datensätze dargestellt. Die zugrunde liegenden Messergebnisse sind zusätzlich im Anhang in den Tabellen A.1 und A.2 nachzulesen. Da die Messergebnisse sowohl bei der Verwendung des Sampson-Fehlers als auch für den 2D-Reprojektionsfehler sehr ähnlich sind, werden in der Abbildung zur besseren Übersicht nur die Kurvenverläufe für den Sampson-Fehler als Kostenfunktion dargestellt. Die Kurven zeigen, dass der normalisierte 8-Punkt-Algorithmus (n8pt) fast ausschließlich den kleinsten Fehler liefert, da in diesem Versuch keine Ausreißer in Form von falschen Punktkorrespondenzen erzeugt wurden. Mit steigendem Rauschen auf den Datensätzen nimmt der Fehler für alle Verfahren zu. Für den ersten Datensatz (Abbildung 3.3(a)), bei dem der Epipol innerhalb des Bildes liegt, ist der Verlauf für LMedS fast identisch mit dem des normalisierten 8-Punkt-Algorithmus. Ebenfalls sehr gute Ergebnisse liefert

der RANSAC-Algorithmus, der bei dem ersten Datensatz bei einer Standardabweichung des gaußverteilten Rauschens von $\sigma = 1$ und $\sigma = 1,5$ sogar einen kleineren Fehler aufweist als der normalisierte 8-Punkt-Algorithmus. Für den zweiten Datensatz (Abbildung 3.3(b)), bei dem sich der Epipol weit außerhalb des Bildbereichs befindet, liegt die Kurve für RANSAC leicht oberhalb der des normalisierten 8-Punkt-Algorithmus. MLESAC und MAPSAC haben beide sehr ähnliche Ergebnisse, allerdings ist der Fehler für beide Datensätze größer als der Fehler, den n8pt, LMedS und RANSAC liefern.

Zusätzlich wurde für den ersten Datensatz noch der Fehler e_2 (Abstand der Epipole) untersucht. Die Ergebnisse sind in Abbildung 3.5(a) (bzw. im Anhang in Tabelle A.3) dargestellt. Hier liefern RANSAC, MLESAC und MAPSAC die kleinsten relativen Fehler, wobei RANSAC ab einer Standardabweichung von $\sigma = 1,5$ am besten abschneidet. Im Gegensatz zum Verlauf des Fehlers e_1 steigt der Fehler e_2 bei n8pt und LMedS ab einer Standardabweichung von $\sigma = 0,5$ sehr stark an. Da bei allen Verfahren der Epipol-Abstand nicht direkt in der zu minimierenden Kostenfunktion enthalten ist und vor allem der n8pt keine Ausreißer eliminiert, könnten schlechte Punktkorrespondenzen diesen Fehler in hohem Maße negativ beeinflussen.

Versuch 2: Im zweiten Versuch wird auf die Koordinaten der beiden Testdatensätze gaußverteiltes Rauschen mit einer konstanten Standardabweichung $\sigma = 0,5$ Pixel gegeben. Ferner wird der Anteil an Ausreißern schrittweise von 5% bis 60% erhöht, indem innerhalb der Datensätze falsche Zuordnungen generiert werden, die über den Bildbereich gleichverteilt sind. Für die vier robusten Ansätze wird jeweils die Fundamentalmatrix entsprechend Versuch 1 ermittelt. Die zu minimierenden Kostenfunktionen basieren ebenfalls zum einen auf dem Sampson-Fehler und zum anderen auf dem 2D-Reprojektionsfehler. Für jeden der robusten Ansätze wird die Fundamentalmatrix 50 mal geschätzt, die Fehler e_1 und e_2 jeweils bestimmt und über die 50 Iterationen gemittelt.

In den Tabellen 3.1 und 3.2 wird dargestellt, inwieweit die robusten Ansätze Ausreißer detektieren bzw. eliminieren. Die erste Zeile in den Tabellen gibt jeweils die Anzahl an tatsächlich erzeugten Ausreißern an. Da die Unterschiede zwischen den beiden verwendeten Kostenfunktionen (Sampson-Fehler und 2D-Reprojektionsfehler) sehr gering sind (siehe Tabelle 3.1), wurde für den zweiten Datensatz lediglich der Sampson-Fehler minimiert. An dieser Stelle soll erwähnt werden, dass die Minimierung des 2D-Reprojektionsfehlers durch die Triangulation zusätzliche Rechenkapazitäten benötigt.

Die Tabellen zeigen, dass RANSAC im Vergleich zum tatsächlichen Anteil an Ausreißern diese am besten detektieren kann. LMedS kann ebenfalls zumindest für den ersten Datensatz (Tabelle 3.1) bis zu einem Anteil an Ausreißern von $20 - 30\%$ diese eliminieren. Ab 30% allerdings weicht die tatsächliche Anzahl an Ausreißern sehr stark von der Anzahl der eliminierten Ausreißer ab. Zudem weisen die hohen Standardabweichungen der

Tabelle 3.1.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern (Datensatz 1, $\sigma = 0,5$ Pixel): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400, Steigerung des Ausreißeranteils von 5% bis 60%, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler und der 2D-Reprojektionsfehler verwendet.

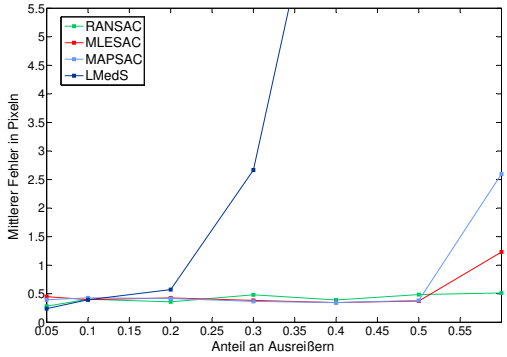
	Mittlere Anzahl an Ausreißern (±StdAbw.)						
Ausreißer	20	40	80	120	160	200	240
RANSAC-Samp	**21,12** (3,48)	40,54 (4,50)	**77,76** (1,79)	**115,52** (2,20)	152,74 (1,78)	192,36 (1,82)	229,60 (1,67)
RANSAC-2Drep	21,64 (4,06)	**39,70** (2,17)	77,42 (1,86)	115,36 (1,57)	**153,02** (1,45)	192,56 (1,77)	**230,08** (1,45)
MLESAC-Samp	150,82 (11,77)	158,90 (10,25)	178,90 (9,36)	198,44 (3,30)	200,58 (0,67)	**200,18** (0,39)	200,16 (0,37)
MLESAC-2Drep	152,78 (14,76)	161,96 (11,13)	179,82 (8,66)	198,10 (4,25)	201,42 (0,88)	200,30 (0,51)	200,68 (0,55)
MAPSAC-Samp	132,10 (9,78)	140,68 (10,78)	157,62 (6,50)	175,72 (4,53)	184,26 (3,47)	197,54 (1,03)	227,96 (40,53)
MAPSAC-2Drep	129,30 (9,12)	138,58 (12,29)	157,78 (9,50)	176,48 (4,62)	184,42 (2,96)	197,50 (1,27)	228,40 (44,33)
LMedS-Samp	28,50 (9,28)	48,30 (10,98)	83,58 (9,08)	106,30 (28,37)	101,50 (53,85)	62,14 (62,24)	26,66 (29,18)
LMedS-2Drep	28,86 (9,36)	46,62 (8,16)	80,50 (6,45)	117,08 (10,26)	149,36 (10,73)	139,42 (38,78)	81,82 (23,78)

Ergebnisse, die über die 50 Iterationen ermittelt wurden, vor allem beim zweiten Datensatz (Tabelle 3.2) darauf hin, dass die Eliminierung innerhalb der Iterationsschritte sehr stark variiert. Dies bedeutet, dass LMedS für einen Anteil an Ausreißern $> 30\%$ im erweiterten Sinne nicht mehr als deterministisch angesehen werden kann. Bei MLESAC und MAPSAC weicht die Anzahl an eliminierten Ausreißern tendentiell sehr stark von der Anzahl an tatsächlich generierten Ausreißern ab, was darauf hindeutet, dass das zugrunde liegende Mischmodell, welches in beiden Ansätzen verwendet wird, nicht hinreichend bestimmt werden konnte.

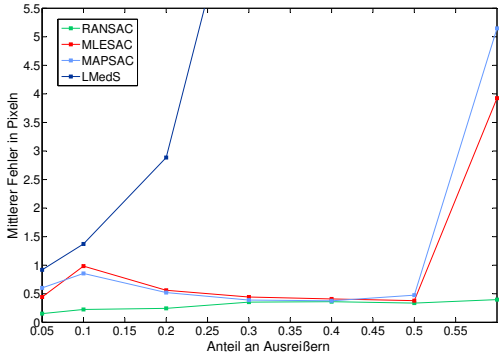
In Abbildung 3.4 sind die mittleren Abstände der Epipolarlinien zu den korrespondierenden Punkten in Abhängigkeit vom Ausreißeranteil für die beiden erzeugten Datensätze dargestellt. Die zugrunde liegenden Messergebnisse sind zusätzlich im Anhang in den Tabellen A.4 und A.5 zu finden. Da die Messergebnisse sowohl bei der Verwendung des

Tabelle 3.2.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern (Datensatz 2, $\sigma = 0,5$ Pixel): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 500, Steigerung des Ausreißeranteils von 5% bis 60%, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler verwendet.

	Mittlere Anzahl an Ausreißern (±StdAbw.)						
Ausreißer	25	50	100	150	200	250	300
RANSAC	**30.46** (8,72)	**55.24** (8,57)	**103.84** (6,79)	**151.74** (5,23)	**198.62** (3,71)	250.36 (4,89)	**298.26** (3,84)
MLESAC	188.24 (24,77)	207.94 (23,01)	231.82 (13,30)	248.42 (3,51)	250.58 (0,81)	**250.08** (0,27)	250.36 (0,56)
MAPSAC	168.66 (21,56)	175.76 (21,30)	203.52 (16,59)	221.32 (7,10)	232.70 (5,21)	253.92 (33,80)	301.60 (65,10)
LMedS	56.36 (32,68)	85.20 (35,68)	121.82 (33,53)	150.06 (40,71)	159.38 (49,80)	105.74 (67,82)	58.40 (45,06)



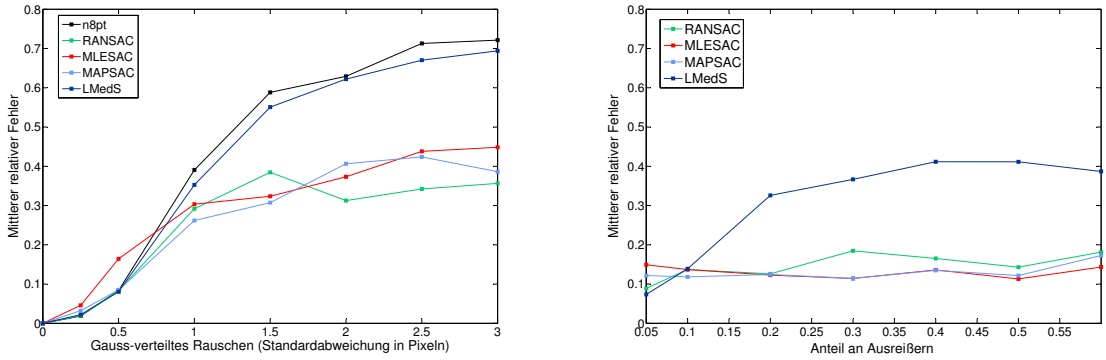
(a) Verlauf der mittleren Abstände zwischen den berechneten Epipolarlinien und den korrekten Bildpunkten (Datensatz 1, Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400)



(b) Verlauf der mittleren Abstände zwischen den berechneten Epipolarlinien und den korrekten Bildpunkten (Datensatz 2, Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 500)

Abbildung 3.4.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern: $\sigma = 0,5$ Pixel (Standardabweichung des gaußverteilten Rauschens), Steigerung des Ausreißeranteils von 5% bis 60%, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler verwendet.

Sampson-Fehlers als auch für den 2D-Reprojektionsfehler wie in Versuch 1 sehr ähnlich sind, werden in der Abbildung wiederum zur besseren Übersicht nur die Kurvenverläufe für den Sampson-Fehler als Kostenfunktion dargestellt. Aufgrund der guten Ergebnisse bei der Eliminierung der Ausreißer, liefert RANSAC unabhängig vom Ausreißeranteil einen annähernd konstanten Verlauf des Fehlers, der weitestgehend unterhalb des Fehlers der anderen robusten Ansätze liegt. Bei LMedS zeigt sich für den ersten Datensatz (Abbil-



(a) Steigerung der Standardabweichung des gaußverteilten Rauschens von 0 bis 3 Pixel

(b) Steigerung des Ausreißeranteils von 5% bis 60% ($\sigma = 0,5$)

Abbildung 3.5.: Verlauf des mittleren relativen Fehlers zwischen dem korrekten Epipol und den geschätzten Epipolen: Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler verwendet.

dung 3.4(a)) ein starker Anstieg des Fehlers bei einem Ausreißeranteil zwischen 20% und 30%, da das Verfahren bei größerer Anzahl an Ausreißern diese nur bedingt eliminieren kann. Für den zweiten Datensatz (Abbildung 3.4(b)) ist ein großer Anstieg des Fehlers bereits ab 10% Ausreißeranteil zu beobachten. Zudem liegt die Kurve für den gesamten Bereich oberhalb der anderen Kurven, d.h. das nichtdeterministische Verhalten bei der Eliminierung von Ausreißern spiegelt sich in dem relativ großen Fehler wieder.

MLESAC und MAPSAC liefern bis zu einem Ausreißeranteil von 50% ähnlich gute Ergebnisse wie RANSAC. Lediglich beim zweiten Datensatz (Abbildung 3.4(b)) ist bei einem Ausreißeranteil von 10% eine größere Abweichung zu beobachten. Der Fehler dort ist etwa um den Faktor 4 größer als bei RANSAC (siehe Tabelle A.5). Obwohl sowohl MLESAC als auch MAPSAC sich zur Eliminierung der Ausreißer nicht oder nur bedingt als geeignet erwiesen haben, lassen sich die guten Ergebnisse der beiden Schätzverfahren im Hinblick auf den geringen Fehler erklären. Beide Verfahren konnten ab einem Ausreißeranteil von 50% diese nicht mehr vollständig eliminieren, weshalb der Fehler sehr stark ansteigt. Bei geringer Anzahl an Ausreißern eliminieren beide Verfahren mehr Ausreißer als tatsächlich durch falsche Zuordnungen erzeugt wurden, d.h. es wurden zusätzlich noch Ausreißer eliminiert, die den gaußverteilten Rauscheinflüssen unterlagen. Diese Tatsache widerspricht nicht dem Grundgedanken robuster Ansätze, da Ausreißer nicht nur falsche sondern auch schlechte Zuordnungen sein können, z. B. Punktkorrespondenzen, deren Koordinaten in größerem Maße durch das Rauschen beeinflusst wurden. D.h. sowohl bei MLESAC als auch bei MAPSAC bildeten die übrigen Punktkorrespondenzen, die nicht eliminiert wurden, einen hinreichend großen Datensatz zur Schätzung einer guten initialen Fundamentalmatrix.

Zusätzlich wurde für den ersten Datensatz ebenfalls noch der Fehler e_2 (Abstand der Epipole) untersucht. Die Ergebnisse sind in Abbildung 3.5(b) (bzw. im Anhang in Tabelle A.6) dargestellt. Hier liefern MLESAC und MAPSAC mit etwa dem gleichen Verlauf die kleinsten relativen Fehler. Der starke Anstieg des Fehlers e_1 ab einem Ausreißeranteil von 50% ist beim Abstand der Epipole e_2 nicht zu beobachten. Der Verlauf des Fehlers bei RANSAC liegt weitestgehend oberhalb dessen von MAPSAC und MLESAC. Lediglich bei kleinen Ausreißeranteilen $< 10\%$ liefert RANSAC bessere Ergebnisse. LMedS hingegen zeigt einen starken Anstieg des Fehlers ab einem Ausreißeranteil von $> 10\%$, so dass auch bei Verwendung des Fehlermaßes e_2 ersichtlich ist, dass dieses Verfahren für eine größere Anzahl an Ausreißern ungeeignet ist.

Versuch 3: Der dritte Versuch entspricht Versuch 2 mit dem Unterschied, dass die konstante Standardabweichung des gaußverteilten Rauschens, welches auf die Testdatensätze addiert wird, auf $\sigma = 2,0$ Pixel erhöht wird. Als Kostenfunktion wird aus den bereits genannten Gründen nur noch der Sampson-Fehler eingesetzt.

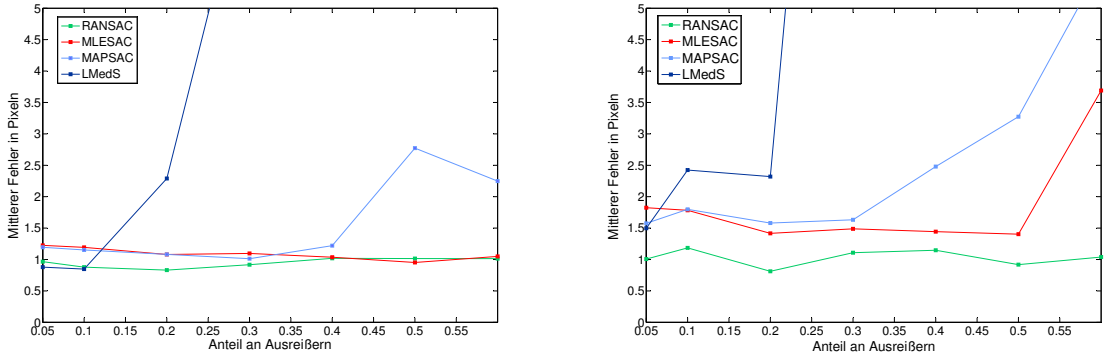
Tabelle 3.3.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern (Datensatz 1, $\sigma = 2,0$ Pixel): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400, Steigerung des Ausreißeranteils von 5% bis 60%, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler verwendet.

	Mittlere Anzahl an Ausreißern (±StdAbw.)						
Ausreißer	20	40	80	120	160	200	240
RANSAC	114,62 (6,56)	130,94 (4,39)	153,30 (4,69)	188,60 (4,14)	210,88 (3,84)	242,40 (2,86)	264,44 (2,78)
MLESAC	187,54 (6,29)	193,72 (6,51)	199,96 (1,35)	200,66 (0,72)	200,50 (0,58)	200,32 (0,55)	200,14 (0,40)
MAPSAC	188,98 (5,18)	196,12 (4,59)	201,74 (3,32)	204,84 (2,68)	210,96 (26,00)	289,08 (67,36)	233,62 (41,82)
LMedS	25,36 (4,89)	40,16 (4,61)	70,30 (18,72)	91,28 (35,12)	97,10 (49,94)	69,40 (55,58)	25,76 (29,18)

In den Tabellen 3.3 und 3.4 wird wiederum dargestellt, inwieweit die robusten Ansätze Ausreißer detektieren bzw. eliminieren. Da durch die Erhöhung des gaußverteilten Rauschens zusätzlich schlechte Punktkorrespondenzen und damit Ausreißer entstehen, sollte die Zahl der eleminierten Ausreißer größer als die Zahl der generierten falschen Zuordnungen (Zeile 1 in den Tabellen) sein. Dies ist nicht für alle Schätzverfahren durchweg zu beobachten. Bei RANSAC trifft diese Aussage für beide Datensätze zu. MAPSAC hingegen eliminiert bei 60% Ausreißeranteil für den ersten Datensatz (Tabelle 3.3) und MLESAC

Tabelle 3.4.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern (Datensatz 2, $\sigma = 2,0$ Pixel): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 500, Steigerung des Ausreißeranteils von 5% bis 60%, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler verwendet.

	Mittlere Anzahl an Ausreißern (±StdAbw.)						
Ausreißer	25	50	100	150	200	250	300
RANSAC	161,68 (9,61)	184,22 (7,65)	200,00 (10,39)	243,28 (6,58)	281,92 (5,66)	308,54 (5,42)	349,56 (4,76)
MLESAC	233,50 (11,36)	238,32 (9,18)	249,00 (3,43)	250,64 (0,66)	250,76 (0,72)	250,22 (0,46)	250,32 (0,51)
MAPSAC	238,66 (8,94)	247,30 (7,14)	251,58 (5,44)	254,74 (4,22)	291,90 (75,68)	351,68 (82,83)	317,72 (66,07)
LMedS	43,46 (18,98)	65,74 (22,39)	106,94 (17,89)	130,72 (35,93)	135,56 (59,23)	107,44 (74,00)	49,66 (41,19)



(a) Verlauf der mittleren Abstände zwischen den berechneten Epipolarlinien und den korrekten Bildpunkten (Datensatz 1, Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400)

(b) Verlauf der mittleren Abstände zwischen den berechneten Epipolarlinien und den korrekten Bildpunkten (Datensatz 2, Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 500)

Abbildung 3.6.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern: $\sigma = 2,0$ Pixel (Standardabweichung des gaußverteilten Rauschens), Steigerung des Ausreißeranteils von 5% bis 60%, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler verwendet.

bei beiden Datensätzen weniger Ausreißer als tatsächlich erzeugt. Zudem nimmt die Anzahl an eliminierten Ausreißern, obwohl der Ausreißeranteil von 5% bis 60% stark ansteigt, bei MLESAC nur geringfügig zu. Für LMedS ist die Anzahl an eliminierten Ausreißern bereits ab etwa 10% kleiner als die tatsächlich erzeugte Anzahl.

Abbildung 3.6 zeigt die mittleren Abstände der Epipolarlinien zu den korrespondierenden Punkten in Abhängigkeit vom Ausreißeranteil für die beiden erzeugten Datensätze.

Im Anhang in den Tabellen A.7 und A.8 sind die zugrunde liegenden Messergebnisse zu finden. Für beide Datensätze liefert RANSAC den kleinsten Fehler, der sich im Vergleich mit Versuch 2 aufgrund des stärkeren Rauschens zwar mehr als verdoppelt hat, aber wiederum annähernd konstant über den gesamten Bereich ist. LMedS liefert nur für den ersten Datensatz (Abbildung 3.6(a)) bis zu einem Ausreißeranteil von 10% gute Ergebnisse. Bei MLESAC ist zumindest für den ersten Datensatz annährend der gleiche Verlauf wie bei RANSAC zu beobachten. Die schlechte Eliminierung der Ausreißer bei 60% spiegelt sich in Abbildung 3.6(a) nicht wider. Für den zweiten Datensatz hingegen (Abbildung 3.6(b)) ist ein starker Anstieg des Fehlers ab 50% Ausreißeranteil festzustellen.

3.5. Kapitelzusammenfassung und Diskussion

In diesem Kapitel wurden Verfahren vorgestellt, mit denen die Epipolargeometrie und damit die Fundamentalmatrix geschätzt werden kann. Diese Verfahren wurden in lineare und nichtlineare Schätzverfahren unterteilt und deren Funktionsweise näher beschrieben. Ferner wurden robuste Ansätze vorgestellt, die die negativen Auswirkungen fehlerhafter Messungen durch Eliminierung von Ausreißern reduzieren. Für die initiale Schätzung der Fundamentalmatrix aus Punktkorrespondenzen sind dabei die sogenannten Monte-Carlo-Methoden besonders geeignet, weshalb mehrere Varianten dieser Ansätze näher untersucht und verglichen wurden. Eine nichtlineare Optimierung im Anschluss an die initiale Schätzung der Fundamentalmatrix wurde hier nicht vorgenommen, da die vorgelegten Ergebnisse lediglich die Robustheit und Genauigkeit der Algorithmen für die Initialisierung aufzeigen sollten.

Die Untersuchung ergab, dass alle diese robusten Verfahren bei geringen Anteilen an Ausreißern für eine Initialisierung einsetzbar sind. Bei Zunahme des Ausreißeranteils zeigte sich, dass LMedS nicht mehr als robust angesehen werden kann. Ferner wurde deutlich, dass RANSAC insgesamt bei allen drei Versuchen gute bzw. die besten Ergebnisse lieferte. Verfahren wie MLESAC und MAPSAC, die zusätzlich die Verteilung der Fehler modellieren und damit eine Erweiterung von RANSAC darstellen, konnten die Erwartungen, die aufgrund ihrer erhöhten Komplexität in sie gelegt wurden, nur zum Teil erfüllen. Vor allem bei einer großen Zahl an Ausreißern lieferten sie sehr schlechte Ergebnisse. Die Ursache für dieses Verhalten ist schwer zu begründen. Ein wesentlicher Faktor liegt sicherlich in den zusätzlichen Parametern, die für das Mischmodell geschätzt werden müssen, und damit in der höheren Komplexität dieser Verfahren. Obwohl die Verfasser dieser Ansätze in [47, 184, 186] fundiert die grundlegende Erweiterung von RANSAC herleiten und damit eine Verbesserung begründen, sind diese nur unzureichend bzw. nicht experimentell miteinander verglichen worden. In [186] z. B. wurden nur die Fehler für die Punktkorrespondenzen berechnet, die nicht eliminiert wurden. Die Modifikation von MAPSAC, wie

in [47] beschrieben, wurde lediglich mit MLESAC und dem ursprünglichen MAPSAC in [184] verglichen, nicht aber mit RANSAC.

Aufgrund der Untersuchung wird im weiteren Verlauf dieser Arbeit der RANSAC-Algorithmus als robustes Schätzverfahren verwendet.

4. 3D-Rekonstruktion aus zwei Ansichten

Nachdem die theoretischen Grundlagen bezüglich Kameramodelle, projektive und Epipolargeometrie bekannt sind, wird nun die 3D-Rekonstruktion aus zwei Ansichten beschrieben, die ebenfalls als Hintergrund für die in Kapitel 5 beschriebene Rekonstruktion aus mehreren Ansichten dienen soll.

4.1. Problemstellung und Überblick über die Methode

Bei der Rekonstruktion einer dreidimensionalen Szene aus den Informationen zweier Bilder, die von unterschiedlicher Perspektive aufgenommen wurden, ist das eigentliche Problem, vor allem diejenigen Informationen auszunutzen, die beide Bilder gemeinsam haben. Mithilfe der Gemeinsamkeiten lassen sich Relationen zwischen den beiden Ansichten ableiten, welche dazu genutzt werden können, die internen und externen Kameraparameter zu schätzen und die 3D-Struktur mit Triangulationsverfahren zu ermitteln.

Die Korrespondenzanalyse zur Bestimmung und Zuordnung von Merkmalsobjekten in beiden Bildern dient als Grundlage für die Ermittlung von projektiven Beziehungen zwischen den Ansichten. Projektive Beziehungen lassen sich über die Epipolargeometrie beschreiben, welche sowohl für die eigentliche 3D-Rekonstruktion als auch für weitere Zuordnung von Merkmalsobjekten verwendet wird. Dabei hängt die Epipolargeometrie nur von den Kameraparametern ab und ist unabhängig von der aufgenommenen Szene. Die Korrespondenzen werden lediglich benötigt, um diese Parameter zu schätzen. Die Genauigkeit dieser Schätzung hängt im großen Maße sowohl von der Fehlerfreiheit der Korrespondenzanalyse als auch von der Lage der Kameras zueinander ab. Auch wenn die Epipolargeometrie nicht direkt abhängig von der aufgenommen Szene ist, hängt das Schätzergebnis zumindest indirekt von der Struktur der Szene ab.

Zusammenfassend stehen drei Kriterien für die Berechnung der Kameraparameter und der 3D-Struktur im Vordergrund, welche die Rekonstruktionsergebnisse signifikant beeinflussen:

- Lage der Kameras zueinander
- Genauigkeit der Korrespondenzanalyse
- Struktur der aufgenommenen Szene

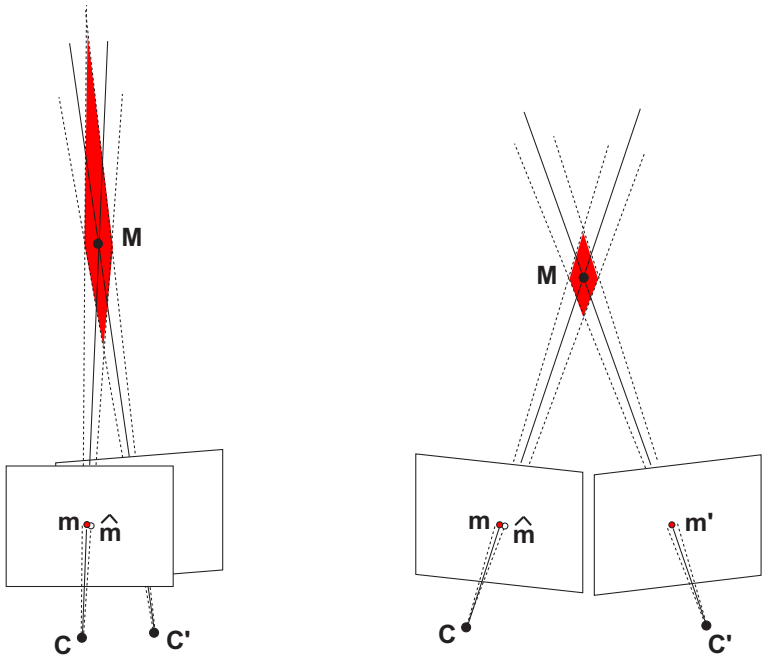


Abbildung 4.1.: Abhängigkeit des zu erwartenden Fehlers bei der Triangulation in Abhängigkeit vom Abstand der Kamerazentren. Der rote Bereich kennzeichnet die Region, in der ein rekonstruierter 3D-Punkt liegen kann, bei gleicher Unsicherheit in der Korrespondenzanalyse. Ein großer Abstand mit gleichzeitiger Veränderung der Ausrichtung der Kameras zueinander verringert den zu erwartenden Fehler bei der Triangulation.

Für den ersten Punkt zeigt Abbildung 4.1 ein Beispiel. Ausgehend von der gleichen Unsicherheit bei der Korrespondenzanalyse von Merkmalspunkten ist der zu erwartende Fehler bei der Triangulation abhängig vom Abstand der Kamerazentren und von der Ausrichtung der Kameras. Ein geringerer Kameraabstand zum Beispiel hat einen größeren Bereich, in dem der zu rekonstruierende 3D-Punkt liegen kann, zur Folge. Mit zunehmendem Kameraabstand erschwert sich allerdings die Korrespondenzanalyse. Zum einen vergrößert sich der Suchbereich, in dem korrespondierende Punkte liegen können, und zum anderen erhöht sich die Wahrscheinlichkeit von Beleuchtungsänderungen bzw. Verdeckungen, welche eine Zuordnung fehleranfälliger bzw. unmöglich machen.

In [187] wurde eine Untersuchung zur optimalen Positionierung und Orientierung zweier Kameras durchgeführt. Es wurde gezeigt, dass bei der Rekonstruktion eines 3D-Objekts der Konvergenz-Winkel der Kameras bezüglich des Objekts im Bereich von 25°-30° liegen sollte. Das Verhältnis der Abstände der Kameras relativ zum Objekt sollte in etwa bei 1,2 liegen. Ferner wurde gezeigt, dass vor allem mit zunehmendem Winkel der Rekonstruktionsfehler stark ansteigt. Als Ursache werden fehlerhafte Zuordnungen bei der Korrespondenzanalyse angegeben.

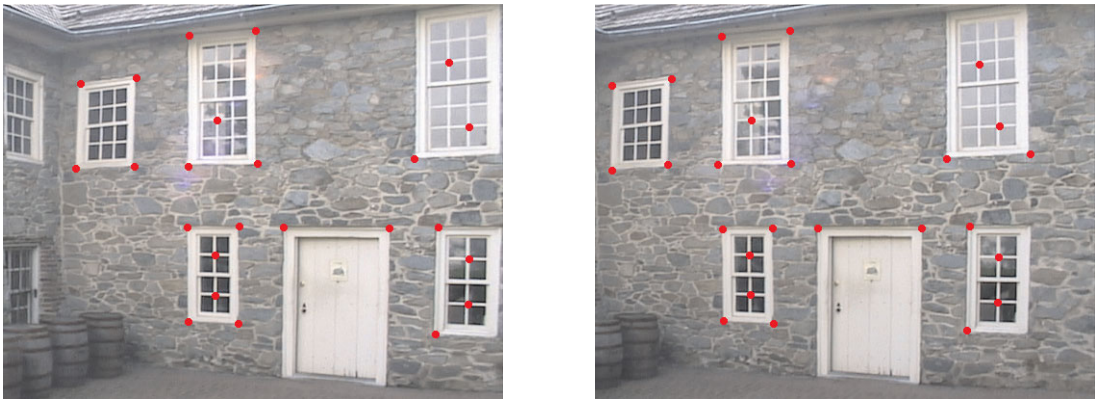


Abbildung 4.2.: Degeneration der 3D-Struktur: Die aufgenommene Szene besteht fast ausschließlich aus einer Ebene.

Das letzte Kriterium, die Struktur der aufgenommenen Szene, spielt dahingehend eine Rolle, dass bei der Degeneration der 3D-Struktur die Kameraparameter nicht eindeutig bestimmt werden können [68,185]. Abbildung 4.2 zeigt dafür ein Beispiel. Die aufgenommene Szene ist eine ebene Hauswand, auf der ausschließlich die Merkmalspunkte liegen. In diesem Fall ist die Epipolargeometrie nicht eindeutig bestimmbar.

Auf Grundlage der zuvor beschriebenen Problemstellung bei der 3D-Rekonstruktion aus zwei Ansichten sind folgende Arbeitsschritte notwendig:

- Korrespondenzanalyse von Merkmalsobjekten
- Robuste Bestimmung der Epipolargeometrie und damit der Fundamentalmatrix
- Schätzung der Kameraparameter und der 3D-Struktur

4.2. Korrespondenzanalyse

In allen Anwendungsgebieten, welche sich mit der Rekonstruktion einer dreidimensionalen Szene beschäftigen, nimmt die Korrespondenzanalyse die wohl wichtigste Rolle ein. Neben vielen anderen Aspekten bei der Rekonstruktion werden Korrespondenzen von Merkmalsobjekten benötigt, um Relationen zwischen zwei oder mehreren Ansichten bestimmen zu können. Dabei stellt die Mehrdeutigkeit dieser Korrespondenzen das größte Problem dar, d.h. Merkmalsobjekte in einer Ansicht können nicht eindeutig den selben Merkmalsobjekten in einer anderen Ansicht zugeordnet werden. Die Fragen, die sich nun stellen, sind: Was sind gute Merkmalsobjekte und wie lassen sich diese in anderen Ansichten finden?

Zwei wesentliche Eigenschaften sollten Merkmalsobjekte aufweisen. Zum einen sollten sie sich von anderen Bildinhalten abheben bzw. hervorstechen und zum anderen sollten sie zwischen verschiedenen Ansichten möglichst unverändert bleiben. Ein Gemälde z. B. hebt sich sehr gut von einer weißen Wand ab und lässt sich in einer anderen Ansicht

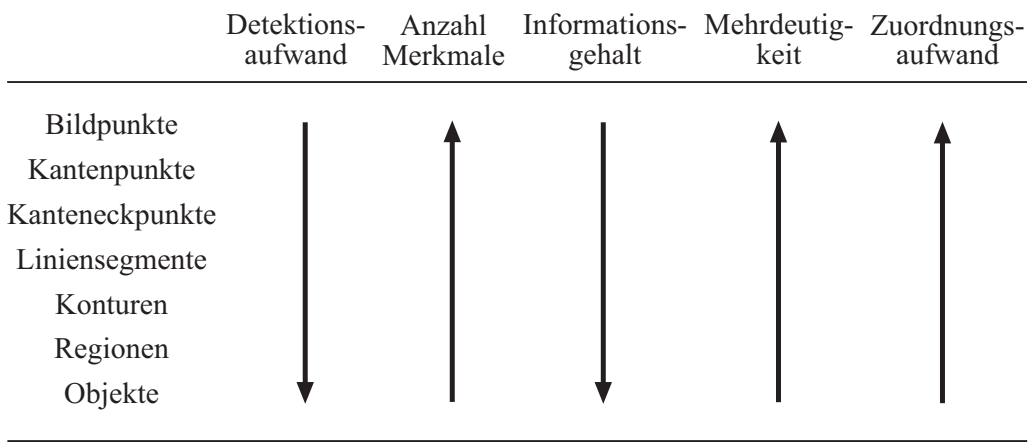


Abbildung 4.3.: Eigenschaften von Merkmalsobjekten

sicherlich leicht wiederfinden, falls es nicht verdeckt wurde oder sich die Lichtverhältnisse drastisch ändern. Ein Mensch, der vor dem Gemälde steht, würde sich z. B. ebenfalls von anderen Bildinhalten abheben. Verändert sich allerdings seine Position, Haltung oder gar die Kleidung, so wird es schwieriger, ihn in einer anderen Ansicht zuzuweisen.

Merkmalsobjekte können Bildpunkte, Kantenpunkte, Eckpunkte, Liniensegmente, Konturen, Regionen oder ganze Objekte sein. In Abbildung 4.3 werden die verschiedenen Merkmalsobjekte mit ihren typischen Eigenschaften dargestellt und untereinander verglichen. Die Pfeile geben jeweils die Zunahme der jeweiligen Eigenschaft an.

Im weiteren Verlauf dieser Arbeit werden Merkmalspunkte verwendet, da diese - vor allem in Hinblick auf Videosequenzen mit geringen Änderungen des Inhalts benachbarter Bilder - geringe Komplexität aufweisen und sich dennoch durch den Einsatz einfacher Techniken zuordnen lassen.

4.2.1. Merkmalspunktdetektor

Eine Vielzahl von Ansätzen zur Detektion von Merkmalspunkten lassen sich in zahlreichen Publikationen finden. Diese Ansätze können in drei Kategorien unterteilt werden: konturbasierte Ansätze, intensitätsbasierte Ansätze und Ansätze, die auf einem parametrischen Modell basieren. Konturbasierte Ansätze extrahieren in einem ersten Schritt Konturen innerhalb des Bildes und suchen dann entlang der Konturen nach den stärksten Krümmungen [4, 29, 124]. Intensitätsbasierte Verfahren arbeiten direkt auf den Grauwert-Bildern und suchen lokal nach Helligkeitsänderungen bzw. hohen Gradienten der Helligkeit [11, 32, 50, 64, 96, 128, 167, 180]. Die modellbasierten Ansätze versuchen ein Modell mit Parametern wie Winkel, Grauwerte, Positionen und Unschärfe an Bildbereiche anzupassen [7, 28, 135, 149]. Obwohl sie sehr präzise sind, sind sie beschränkt auf eine spezielle Art

von Merkmalspunkten, z. B. *l-förmige* Eckpunkte wie in Abbildung 4.4 dargestellt. Ein guter Überblick über weitere Merkmalspunktdetektoren und umfangreiche Evaluierungen sind in Deriche und Giraudon [30] und Schmid et al. [158] zu finden.

In dieser Arbeit wurde der wohl populärste Merkmalspunktdetektor, der Harris Detektor [64], verwendet, welcher im Weiteren näher beschrieben werden soll.

Harris Detektor

Der Harris Detektor zählt zu den intensitätsbasierten Merkmalspunktdetektoren und stellt eine Erweiterung des *Moravec Detektors* [128] dar. Dabei wird der Gradient der Helligkeitswerte durch die partielle Ableitung erster Ordnung sowohl in vertikaler als auch horizontaler Richtung ermittelt:

$$I_x = \frac{\partial I}{\partial x} = I \otimes (-1, 0, 1) \tag{4.1}$$

$$I_y = \frac{\partial I}{\partial y} = I \otimes (-1, 0, 1)^T . \tag{4.2}$$

Die lokale Autokorrelationsfunktion für kleine Fensterversätze wird wie folgt definiert:

$$E(x, y) = Ax^2 + 2Cxy + By^2 \tag{4.3}$$

mit

$$\begin{aligned} A &= I_x^2 \otimes \mathcal{W} \\ B &= I_y^2 \otimes \mathcal{W} \\ C &= (I_x I_y) \otimes \mathcal{W} \end{aligned}$$

und $\mathcal{W}$ als Fenster um einen Merkmalspunkt. Gl. 4.3 kann in Matrixschreibweise angegeben werden mit

$$E(x, y) = (x, y) M (x, y)^T, \tag{4.4}$$

wobei M eine symmetrische 2×2-Matrix (Gradientenmatrix)

$$M = \begin{bmatrix} A & C \\ C & B \end{bmatrix} \tag{4.5}$$

ist. Seien α und β die Eigenwerte der Matrix M, α und β stellen dann ein Maß für die Krümmung der lokalen Autokorrelationsfunktion dar. Große Eigenwerte bedeuten eine starke Krümmung der Autokorrelationsfunktion und weisen damit auf Merkmalspunkte

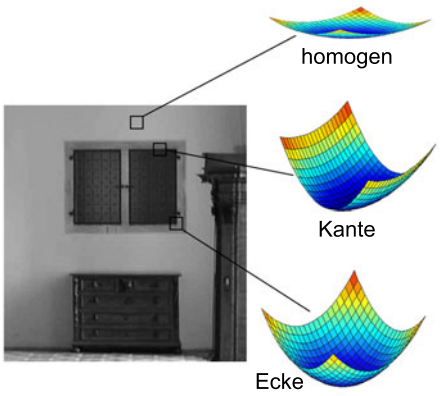


Abbildung 4.4.: Vergleich verschiedener Autokorrelationsverläufe für homogene Regionen, Kanten und Eckpunkte.

Abbildung 4.5.: Verteilung von Merkmalspunkten durch Aufteilung des Bildes in unabhängige Bereiche.

hin, welche große Helligkeitsunterschiede in vertikaler und horizontaler Richtung besitzen. Diese Merkmalspunkte sind dann vor allem Eckpunkte (engl. *corner*) in einem Grauwertbild. Durch die Verwendung von Gradienten wird der Merkmalspunktdetektor invariant gegenüber Beleuchtungsänderungen. Ferner führt die Verwendung von Eigenwerten der Gradientenmatrix zu einer Invarianz gegenüber Drehungen des Bildinhalts.

In Abbildung 4.4 ist ein Beispiel für drei Fälle der lokalen Autokorrelationsfunktion dargestellt. Homogene Regionen, in denen die Intensitäten annähernd konstant sind, weisen nur eine schwache Krümmung in beide Richtungen auf. Für Bildkanten sind Helligkeitsunterschiede lediglich orthogonal zum Verlauf der Kante vorhanden. Die Autokorrelationsfunktion besitzt somit nur in einer Richtung eine stärkere Krümmung. Im Falle von Eckpunkten, bei denen die Intensitäten in beiden Richtungen variieren, ist der Verlauf der Krümmung sowohl in vertikaler als auch in horizontaler Richtung stark ausgeprägt.

Anstelle der Eigenwertzerlegung, die für die Berechnung von α und β erforderlich wäre, wird in [64] ein Maß für die Krümmung der lokalen Autokorrelationsfunktion über die Spur und die Determinate der Matrix M definiert:

$$R = \mathbf{det}(M) - k \cdot \mathbf{spur}(M)^2. \tag{4.6}$$

Für den Parameter k wird von Harris ein Wert von $0,04$ vorgeschlagen.

Mit $spur(M) = \alpha + \beta = A + B$ und $\mathbf{det}(M) = \alpha\beta = AB - C^2$ lassen sich die drei Fälle für die Krümmung der Autokorrelationsfunktion über Gl. 4.6 wie folgt beschreiben:

- homogene Region: $\alpha, \beta \rightarrow 0$. Damit wird R klein bzw. $R \rightarrow 0$.
- Kante: $\alpha \rightarrow 0$ oder $\beta \rightarrow 0$. Bei einer Kante ist einer der beiden Eigenwerte klein bzw. wird Null. Daraus folgt, dass $\mathbf{det}(M) \rightarrow 0$ und somit R negativ wird.

- Eckpunkte: Für große Eigenwerte α, β nimmt R positive Werte an. R wächst proportional mit der Größe der Eigenwerte und damit mit der Krümmung der lokalen Autokorrelationsfunktion.

Durch die Festlegung eines Schwellwertes für R lässt sich die Anzahl der Merkmalspunkte regulieren. Ein hoher Schwellwert führt zu einer Reduzierung der Extraktion von Merkmalspunkten, wobei die Qualität der Merkmalspunkte, welche sich durch die Stärke der Krümmung der lokalen Autokorrelationsfunktion ausdrückt, zunimmt.

Häufig treten Merkmalspunkte in ein paar wenigen lokalen Bereichen des Bildes auf. In vielen Anwendungsszenarien von Merkmalspunktdetektoren, z. B. bei der Rekonstruktion der 3D-Struktur, kann eine gleichmäßige Verteilung der Merkmalspunkte über alle Bildbereiche aber sinnvoll sein, um mögliche Degenerationen in anschließenden Parameterschätzverfahren zu vermeiden. In Abbildung 4.5 ist ein Beispiel für die möglichst gleichmäßige Verteilung von Merkmalspunkten im Bild gegeben. Das Bild wurde zuvor in vier unabhängige Bereiche aufgeteilt, in denen separat mit unterschiedlichen Schwellwerten für R Merkmalspunkte detektiert wurden.

4.2.2. Suche nach eindeutigen Zuordnungen

Geht man nun davon aus, dass in zwei Bildern, die von unterschiedlicher Perspektive aufgenommen wurden, Merkmalspunkte detektiert wurden, ist der nächste Schritt zur 3D-Rekonstruktion, Zuordnungen dieser Punkte (engl. *feature matching*) in beiden Ansichten zu finden (d.h. die eigentliche Korrespondenzanalyse).

Eines der größten Probleme in Anwendungsbereichen der Computer Vision ist die eindeutige Zuordnung solcher Merkmalspunkte in zwei oder mehreren Ansichten. Die Forschungsarbeiten in diesem Feld haben eine lang andauernde Tradition und dennoch sind die Probleme der Mehrdeutigkeit von Zuordnungen nicht vollends gelöst. Dem entsprechend lässt sich eine Vielzahl von Publikationen zu diesem Thema in der Literatur finden [5, 29, 35, 37, 51, 52, 71, 90, 107, 142, 147, 156, 190, 201]. Die meisten Verfahren verwenden zusätzliche Randbedingungen, welche die Mehrdeutigkeiten von Korrespondenzen reduzieren, z. B. die Einzigartigkeitsbedingung, die Epipolarbedingung, die Disparitätsbedingung oder die Nachbarschaftsbedingung, auf die im Weiteren noch eingegangen wird.

Bei der Korrespondenzanalyse wird die Zuordnung von Merkmalspunkten anhand der Ähnlichkeit des umliegenden Bildbereichs, dem sogenannten Korrelationsfenster, in beiden Ansichten ermittelt. Die am häufigsten verwendeten Ähnlichkeitsmessungen beruhen auf der *Summe der absoluten Abweichungen* (engl. *sum of absolute differences*, SAD), der *Summe der quadratischen Abweichungen* (engl. *sum of squared differences*, SSD) oder der *normalisierten Kreuzkorrelation* (engl. *normalized cross-correlation*, NCC). Trujillo und Izquierdo [190]

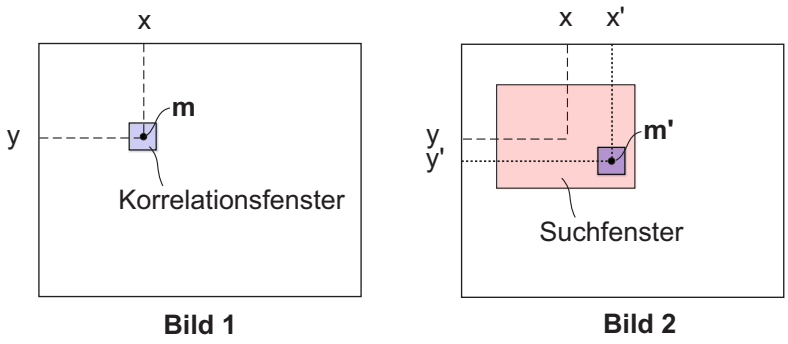


Abbildung 4.6.: Zuordnung durch Korrelation innerhalb eines Suchbereichs.

stellten einen robusten Zuordnungsansatz vor, welcher auf einer leicht abweichenden Version der NCC beruht, der sogenannten *Median-Korrelation* (engl. *median correlation*, MC), die robuster gegen Bildrauschen und Tiefensprünge ist. In Scharstein und Szeliski [156] und Jones [90] wird ein guter Überblick über robuste Ansätze bei der Korrespondenzanalyse gegeben.

Da korrelationsbasierte Zuordnungen resistenter gegen Rotationen und Helligkeitsänderungen sind als Ähnlichkeitsmessungen, die nur auf Differenzen der Grauwerte basieren wie SAD oder SSD, wird im Weiteren vorerst die NCC als Zuordungsmaß verwendet. Nach der Detektion von Merkmalspunkten in zwei Bildern wird für jeden dieser Punkte im ersten Bild innerhalb eines Suchbereichs im zweiten Bild nach dem korrespondierenden Punkt gesucht. Die Größe dieses Suchbereichs hängt von den zu erwartenden maximalen Disparitäten in horizontaler ($\pm d_x$) und vertikaler Richtung ($\pm d_y$) ab (Disparitätsbedingung). Allgemein wird für das Suchfenster eine rechteckförmige Abmessung mit den Maßen $(2d_x + 1) \times (2d_y + 1)$ verwendet, dessen Zentrum die gleichen Koordinaten wie die des zu suchenden Punktes besitzt. Für alle detektierten Merkmalspunkte, die innerhalb dieses Suchbereichs im zweiten Bild liegen, werden die NCC-Werte über die sogenannten Korrelationsfenster ermittelt. Korrelationsfenster haben ebenfalls eine rechteckförmige Abmessung der Größe $(2n + 1) \times (2m + 1)$ mit dem Merkmalspunkt als Zentrum. Ein Beispiel mit Suchfenster und Korrelationsfenstern ist in Abbildung 4.6 gegeben.

Der NCC-Wert zwischen zwei Korrelationsfenstern aus beiden Bildern wird über die Gleichung

$$\rho\left(\mathbf{m}, \mathbf{m}'\right) = \frac{\sum_{i=-n}^{n} \sum_{i=-m}^{m} \left[I_1(x+i, y+j) - \overline{I_1(x,y)}\right] \cdot \left[I_2(x'+i, y'+j) - \overline{I_2(x',y')}\right]}{(2n+1) \cdot (2m+1) \sqrt{\sigma^2(I_1) \cdot \sigma^2(I_2)}} \tag{4.7}$$

berechnet. $\overline{I_k(x,y)} = \frac{\sum_{i=-n}^{n}\sum_{i=-m}^{m} I_k(x+i,y+j)}{(2n+1)\cdot(2m+1)}$ in Gl. 4.7 gibt den mittleren Helligkeitswert und $\sigma(I_k)$ die Standardabweichung des Helligkeitswertes im Korrelationsfenster an, wobei

$$\sigma(I_k) = \sqrt{\frac{\sum_{i=-n}^{n}\sum_{i=-m}^{m} I_k^2(x+i,y+j)}{(2n+1)\cdot(2m+1)} - \overline{I_k(x,y)}} \tag{4.8}$$

ist.

Der NCC-Wert nimmt durch die Normalisierung nur Werte aus dem Bereich zwischen -1 und 1 an. Ein Wert von -1 heißt, dass die beiden Korrelationsfenster komplementär zueinander sind, und ein Wert von 1 bedeutet, dass beide Korrelationsfenster identisch sind. Eine einfache Zuordnung könnte demzufolge getroffen werden, indem diejenigen Merkmalspunkte mit dem höchsten NCC-Wert als Korrespondenzen ausgewählt werden (engl. *winner-take-all*). In der Praxis würde solch eine einfache Zuordnung allerdings zu zahlreichen falschen Korrespondenzen führen, da Bildrauschen, Helligkeitsänderungen, Ver- und Aufdeckungen nicht auszuschließen sind. Gängige Verfahren [201] bestimmen eine Vorauswahl an möglichen Zuordnungen, indem ein Schwellwert τ für die NCC festgelegt wird. Liegt der NCC-Wert oberhalb dieser Schwelle, so gelten die beiden Merkmalspunkte als Kandidaten für eine mögliche Korrespondenz. Für jeden Punkt in Bild 1 existiert somit eine Menge $\mathcal{M}$ an Kandidaten im Bild 2, die für eine Zuordnung in Frage kommen können. Diese Menge kann, z. B. im Falle einer Verdeckung, auch eine leere Menge sein, d.h. eine Zuordnung des Punktes in Bild 1 zu einem Punkt im Bild 2 wäre nicht möglich.

Nachbarschaftsbedingung

Da die Definition einer Schwelle für den NCC-Wert nicht für eine eindeutige Zuordnung ausreicht, müssen weitere Kriterien herangezogen werden, welche die Mehrdeutigkeit weiter einschränken. Ein äußerst starkes Kriterium für eine eindeutigere Zuordnung ist die Nachbarschaftsbedingung (engl. *neighborhood constraint*). Dabei wird für jeden Kandidaten die Unterstützung benachbarter Punktzuordnungen herangezogen, denn bei einer korrekten Zuordnung eines Merkmalspunktes ist zu erwarten, dass in der umliegenden Nachbarschaft ebenfalls korrekte Zuordnungen liegen. Ist dies nicht der Fall, so gilt die Zuordnung als „unglaubwürdig“ und wäre mit großer Wahrscheinlichkeit eine Falsch-Zuordnung. Die folgenden Ausführungen entstammen weitestgehend aus der Publikation von Zhang et al. [201].

Betrachtet man einen Kandidaten für eine Zuordnung $\{\mathbf{m}_i \leftrightarrow \mathbf{m}'_j\}$, wobei $\mathbf{m}_i$ ein Merkmalspunkt im ersten Bild und $\mathbf{m}'_j$ der zugeordnete Merkmalspunkt im zweiten Bild ist, und definiert man die Nachbarschaften um diese beiden Punkte in beiden Bildern $\mathcal{N}(\mathbf{m}_i)$

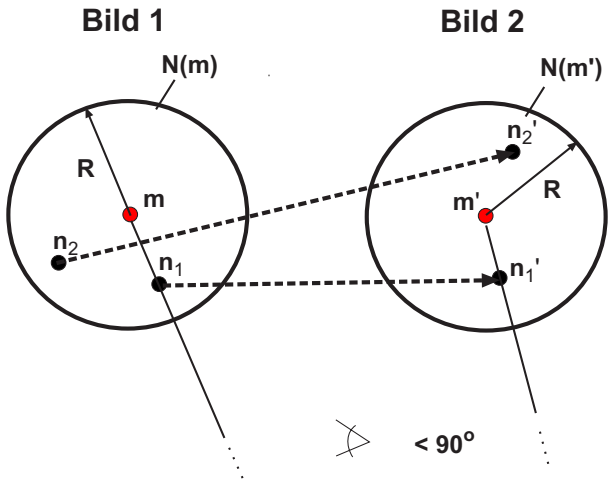


Abbildung 4.7.: Nachbarschaftsbedingungen: Zuordnung mithilfe benachbarter Merkmalspunkte.

und $\mathcal{N}(\mathbf{m}'_j)$ mit einem kreisförmigen Radius R_N, so bilden beide Punkte eine hinreichend gute Zuordnung, wenn eine Vielzahl von Zuordnungen $\{\mathbf{n}_k \leftrightarrow \mathbf{n}'_l\}$ in der Nachbarschaft existiert, wobei $\mathbf{n}_k \in \mathcal{N}(\mathbf{m}_i)$ und $\mathbf{n}'_l \in \mathcal{N}(\mathbf{m}'_j)$ ist. Zudem sollten diese Nachbarschaftszuordnungen in beiden Bildern eine ähnliche relative Position zu dem entsprechenden Kandidaten aufweisen, d.h. die Position von $\mathbf{n}_k$ relativ zu $\mathbf{m}_i$ ist in etwa gleich der Position von $\mathbf{n}'_l$ relativ zu $\mathbf{m}'_j$.

Beispiel: Abbildung 4.7 zeigt dazu ein Beispiel. Dargestellt wird ein Kandidat für eine Zuordnung $\{\mathbf{m} \leftrightarrow \mathbf{m}'\}$ und dessen Nachbarschaft. Die Zuordnungen $\{\mathbf{n}_1 \leftrightarrow \mathbf{n}'_1\}$ und $\{\mathbf{n}_2 \leftrightarrow \mathbf{n}'_2\}$ liegen innerhalb der Nachbarschaft und könnten demnach die Zuordnung des Kandidaten unterstützen. Allerdings ist in diesem Beispiel lediglich die Zuordnung $\{\mathbf{n}_1 \leftrightarrow \mathbf{n}'_1\}$ hilfreich, da nur diese die Bedingung der ähnlichen Position relativ zu dem entsprechenden Kandidaten erfüllt. Die Nachbarschaftszuordnung $\{\mathbf{n}_2 \leftrightarrow \mathbf{n}'_2\}$ besitzt zwar in etwa die gleiche Euklidische Distanz zu dem entsprechenden Kandidaten, der Winkel Θ zwischen den Geraden durch die Punkte $\{\mathbf{m}, \mathbf{n}_2\}$ und die Punkte $\{\mathbf{m}', \mathbf{n}'_2\}$ ist aber zu groß ($\Theta > 90°$), als dass diese Zuordnung als Unterstützung in Frage kommen könnte.

Stärke der Zuordnung. Im Folgenden soll nun ein Maß für die Stärke einer Zuordnung definiert werden. Dazu werden die Bedingungen für die Unterstützung des Kandidaten $\{\mathbf{m}_i \leftrightarrow \mathbf{m}'_j\}$ durch die Nachbarschaft noch einmal zusammengefasst:

1. Die Zuordnung $\{\mathbf{n}_k \leftrightarrow \mathbf{n}'_l\}$, welche den Kandidaten unterstützen soll, muss innerhalb eines definierten Bereichs um den Kandidaten mit dem Radius R_N liegen, d.h.

$\mathbf{n}_k \in \mathcal{N}(\mathbf{m}_i)$ und $\mathbf{n}'_l \in \mathcal{N}(\mathbf{m}'_j)$. Für R_N wird in der Regel als Wert ein Achtel der Bildbreite verwendet [201].

2. Der Winkel zwischen den Geraden $\overline{\mathbf{m}_i\mathbf{n}_k}$ und $\overline{\mathbf{m}'_j\mathbf{n}'_l}$ muss kleiner sein als Θ (mit $\Theta = 90°$).

Sind diese Bedingungen erfüllt, lässt sich ein Maß für die Stärke der Zuordnung definieren mit

$$S_Z\left(\mathbf{m}_i, \mathbf{m}'_j\right) = \sum_{\mathbf{n}_k \in \mathcal{N}(\mathbf{m}_i)} \sum_{\mathbf{n}'_l \in \mathcal{N}(\mathbf{m}'_j)} \max_{\{\mathbf{n}_k \leftrightarrow \mathbf{n}'_l\}} \frac{\rho(\mathbf{m}_i, \mathbf{m}'_j) \cdot \rho(\mathbf{n}_k, \mathbf{n}'_l) \cdot \delta(\mathbf{m}_i, \mathbf{m}'_j; \mathbf{n}_k, \mathbf{n}'_l)}{1 + dist(\mathbf{m}_i, \mathbf{m}'_j; \mathbf{n}_k, \mathbf{n}'_l)}. \quad (4.9)$$

Dabei ist

$$dist(\mathbf{m}_i, \mathbf{m}'_j; \mathbf{n}_k, \mathbf{n}'_l) = \left[d(\mathbf{m}_i, \mathbf{n}_k) + d(\mathbf{m}'_j, \mathbf{n}'_l)\right] / 2 \quad (4.10)$$

der mittlere symmetrische Abstand des Kandidaten zu einem Nachbarn in beiden Bildern. $d(\mathbf{m}, \mathbf{n})$ entspricht der Euklidischen Distanz zwischen $\mathbf{m}$ und $\mathbf{n}$. $\delta(\mathbf{m}_i, \mathbf{m}'_j; \mathbf{n}_k, \mathbf{n}'_l)$ ist eine Exponentialfunktion, abhängig von der relativen Differenz r der Abstände in beiden Bildern:

$$\delta(\mathbf{m}_i, \mathbf{m}'_j; \mathbf{n}_k, \mathbf{n}'_l) = \begin{cases} e^{-r/\epsilon_r} & wenn\ r < \epsilon_r \\ 0 & sonst \end{cases}, \quad (4.11)$$

mit

$$r = \frac{\left|d(\mathbf{m}_i, \mathbf{n}_k) - d(\mathbf{m}'_j, \mathbf{n}'_l)\right|}{dist(\mathbf{m}_i, \mathbf{m}'_j; \mathbf{n}_k, \mathbf{n}'_l)} \quad (4.12)$$

und ϵ_r als Schwellwert für die relative Differenz r der Abstände. In [201] wird für ϵ_r ein Wert von $0,3$ vorgeschlagen, welches einer relativen Abweichung vom mittleren Abstand von 15% entspricht. Die Verwendung eines relativen Maßes für die Differenz ist dahingehend sinnvoll, da damit die Berechnungen bei der Korrespondenzanalyse unabhängig von der Bildauflösung sind. Ferner steigt mit zunehmendem Abstand der Nachbarzuordnungen zum Kandidaten in der Regel auch die absolute Differenz dieser Abstände aufgrund perspektivischer Verzerrungen in beiden Bildern.

Für die Stärke der Zuordnung aus Gl. 4.9 ergibt sich durch die Summation und die Verwendung der mittleren Abstände der Nachbarschaftszuordnungen im Nenner ein höheres Maß, je mehr Nachbarschaftszuordnungen vorhanden sind und je dichter diese an dem eigentlichen Kandidaten liegen. Da auch unter den Nachbarschaftspunkten mehrere Zuordnungen möglich sind, wird durch den Maximum-Operator diejenige Zuordnung gewählt, welche die geringste relative Differenz r der Abstände aufweist. Zusätzlich reguliert die Exponentialfunktion im Zähler von Gl. 4.9 durch den negativen Exponenten die Größe jedes Summanden, d.h. die ausgewählten Nachbarschaftszuordnungen tragen umso mehr zu S_Z bei, je kleiner deren relative Differenz r ist.

Relaxation. Durch die Nachbarschaftsunterstützung, wie sie bislang eingeführt wurde, lässt sich die Glaubwürdigkeit einer Zuordnung gegenüber der einfachen NCC-basierten Zuordnung stark erhöhen. Dennoch lassen sich falsche Korrespondenzen, verursacht durch eine Zuordnung ausschließlich aufgrund des größten Wertes für S_Z, noch weiter reduzieren, indem z. B. die Bedingung der Einzigartigkeit einer Korrespondenz (engl. *uniqueness constraint*) verwendet wird. Dies hat insofern eine wesentliche Bedeutung, dass Nachbarschaftszuordnungen $\{\mathbf{n}_k \leftrightarrow \mathbf{n}'_l\}$, die für die Berechnung von S_Z eines Kandidaten $\{\mathbf{m}_i \leftrightarrow \mathbf{m}'_j\}$ herangezogen wurden, vielleicht dann, wenn sie selbst Kandidaten sind, für die S_Z bestimmt wird, gar nicht zugeordnet werden.

Beispiel: Der Merkmalspunkt $\mathbf{m}_1$ in Bild 1 besitzt zwei mögliche Punktkorrespondenzen $\mathbf{m}'_1 \in \mathcal{M}(\mathbf{m}_1)$ und $\mathbf{m}'_6 \in \mathcal{M}(\mathbf{m}_1)$ in Bild 2, da der NCC-Wert oberhalb der Schwelle τ liegt. Berechnet werden die Stärke der Zuordnung $S_Z(\mathbf{m}_1, \mathbf{m}'_1)$ durch Unterstützung der Nachbarschaftszuordnungen $\{\mathbf{n}_2, \mathbf{n}'_2\}$ und $\{\mathbf{n}_3, \mathbf{n}'_3\}$ und $S_Z(\mathbf{m}_1, \mathbf{m}'_6)$ durch Unterstützung der Nachbarschaftszuordnungen $\{\mathbf{n}_3, \mathbf{n}'_4\}$ und $\{\mathbf{n}_5, \mathbf{n}'_5\}$. Sei nun $S_Z(\mathbf{m}_1, \mathbf{m}'_6) > S_Z(\mathbf{m}_1, \mathbf{m}'_1)$, dann würde der Kandidat $\{\mathbf{m}_1, \mathbf{m}'_6\}$ als Korrespondenz bestimmt werden. Wird nun der Merkmalspunkt $\mathbf{m}_3$, der zuvor noch als Unterstützung $\mathbf{n}_3$ der Zuordnungen $\{\mathbf{m}_1, \mathbf{m}'_1\}$ bzw. $\{\mathbf{m}_1, \mathbf{m}'_6\}$ verwendet wurde, als Kandidat mit den möglichen Zuordnungen $\mathbf{m}'_3 \in \mathcal{M}(\mathbf{m}_3)$ und $\mathbf{m}'_4 \in \mathcal{M}(\mathbf{m}_3)$ überprüft, so würde die Zuordnung $\{\mathbf{m}_3, \mathbf{m}'_4\}$ eliminiert werden, wenn $S_Z(\mathbf{m}_3, \mathbf{m}'_3) > S_Z(\mathbf{m}_3, \mathbf{m}'_4)$. Da aber $\{\mathbf{n}_3, \mathbf{n}'_4\}$ zuvor als Nachbarschaftszuordnung bei der Berechnung von $S_Z(\mathbf{m}_1, \mathbf{m}'_6)$ verwendet wurde und ggf. ausschlaggebend für die Zuordnung $\{\mathbf{m}_1, \mathbf{m}'_6\}$ war, besteht eine gegenseitige Beeinflussung der Ergebnisse bei der Bestimmung der Zuordnungen.

In [201] wird eine Erweiterung der *winner-take-all*-Strategie vorgestellt, welche auf einem iterativen Prozess basiert, bei dem nicht alle Zuordnungen mit dem größten Wert für S_Z ausgewählt werden. Dazu wird ein weiteres Maß eingeführt, welches die Glaubwürdigkeit von Zuordnungen G_Z bewertet, indem sowohl die maximale Stärke einer Zuordnung als auch die zweitbeste mögliche Zuordnung eines Kandidaten berücksichtigt wird:

$$G_z = 1 - \frac{S_Z^{(2)}}{S_Z^{(1)}}. \tag{4.13}$$

Zuordnungen, deren Stärke nach Gl. 4.9 zu unterschiedlichen Kandidaten ähnlich groß ist, sollen damit als eher unglaubwürdig bewertet werden. Gl. 4.13 geht gegen 0 (unglaubwürdig), wenn zwei Zuordnungen eines Merkmalspunktes in etwa die gleiche Stärke erzielen, und gegen 1 (glaubwürdig), wenn die zweite mögliche Zuordnung eine geringe Stärke aufweist bzw. keine zweite Zuordnung vorhanden ist. Die Auswahl geeigneter Korrespondenzen wird durch Wiederholung der folgenden Rechenschritte vorgenommen:

1. Für alle möglichen Zuordnungen wird deren Stärke S_Z ermittelt.
2. Für jeden Merkmalspunkt wird die Glaubwürdigkeit G_Z berechnet.
3. Alle Merkmalspunkte aus beiden Bildern werden nach Stärke und Glaubwürdigkeit in absteigender Reihenfolge sortiert. Ein Merkmalspunkt, dessen Stärke und Glaubwürdigkeit zu den obersten q Prozent gehört, wird als *Gewinner* deklariert.
4. Zuordnungen, bei denen beide Merkmalspunkte als Gewinner deklariert wurden, aber zumindest einer dieser Punkte eine größere Stärke der Zuordnung mit einem anderen Merkmalspunkt aufweist, werden gelöscht.

Für q sollte ein Wert zwischen $50\% < q < 100\%$ verwendet werden. Ist z. B. $q \leq 50\%$, so besteht unter Umständen die Möglichkeit, dass keiner der Merkmalspunkte als Gewinner deklariert wird und somit auch keine Zuordnung aussortiert wird. Für $q = 100\%$ würden dagegen alle Merkmalspunkte als Gewinner deklariert werden. In diesem Fall wäre das Verfahren wieder eine *winner-take-all*-Strategie. Sinnvolle Werte für q liegen zwischen 60% und 70%.

Treten keine Veränderungen bei der Wiederholung der Rechenschritte mehr auf, d.h. keine Zuordnungen werden aussortiert, so wird die Iteration abgebrochen. In einem letzten Schritt werden diejenigen Zuordnungen ausgewählt, bei denen beide Merkmalspunkte als Gewinner deklariert wurden und deren Stärke der Zuordnung der maximalen Stärke beider Merkmalspunkte entspricht. Das gesamte Verfahren zur Auswahl von Zuordnungen ist in Algorithmus 3 zusammengefasst.

Abbildung 4.8 zeigt ein Beispiel für die Zuordnung von 400 detektierten Merkmalspunkten in dem Bildpaar *Halle* mit Hilfe der Nachbarschaftsbedingung.

Eine zusätzliche Möglichkeit, die Mehrdeutigkeit noch weiter zu reduzieren, ist die Epipolar-Bedingung, welche in Abschnitt 4.3 im Allgemeinen und in Abschnitt 4.3.3 im Speziellen vorgestellt wird.

4.3. Schätzung der Epipolargeometrie

Nachdem Punktkorrespondenzen $\{\mathbf{m}_i \leftrightarrow \mathbf{m}'_i\}$ in beiden Bildern ermittelt wurden, sollen im weiteren Verlauf die 3D-Koordinaten dieser Punkte und die Projektionsmatrizen bestimmt werden, so dass gilt:

$$\mathbf{m}_i = P\mathbf{M}_i \tag{4.14}$$

$$\mathbf{m}'_i = P'\mathbf{M}_i, \tag{4.15}$$

wobei P und P' die Kameraprojektionsmatrizen, welche sich über die Fundamentalmatrix berechnen lassen, für beide Kameraansichten darstellen. Ziel ist es nun, mit Hilfe der

Ziel: Auswahl eindeutiger Zuordnungen von Merkmalspunkten zwischen zwei Bildern.

- Berechnung der NCC-Werte zwischen allen Merkmalspunkten in beiden Bildern, die innerhalb eines Suchbereichs von $(2d_x + 1) \times (2d_y + 1)$ liegen (Disparitätsbedingung), mit Gl. 4.7.
- Auswahl der Zuordnungen, dessen NCC-Wert größer als der Schwellwert τ ist ($\tau = 0,8$).
  Wiederhole bis keine Änderung mehr auftritt (Nachbarschaftsbedingung)
  1. Berechnung der Stärke S_Z aller möglicher Zuordnungen mit Gl. 4.9.
  2. Bestimmung der Glaubwürdigkeit G_Z für alle Merkmalspunkte mit Gl. 4.13.
  3. Deklaration von Merkmalspunkten als Gewinner, dessen Stärke und Glaubwürdigkeit zu den obersten q Prozent gehören ($q = 60\%$).
  4. Löschen von Zuordnungen, bei denen beide Merkmalspunkte als Gewinner deklariert wurden, aber zumindest einer dieser Punkte eine größere Stärke der Zuordnung mit einem anderen Merkmalspunkt aufweist.

  Ende der Schleife
- Auswahl der Zuordnungen, bei denen beide Merkmalspunkte als Gewinner deklariert wurden und deren Stärke der Zuordnung der maximalen Stärke beider Merkmalspunkte entspricht.

Ergebnis: Merkmalspunkte mit eindeutigen Zuordnungen in beiden Bildern.

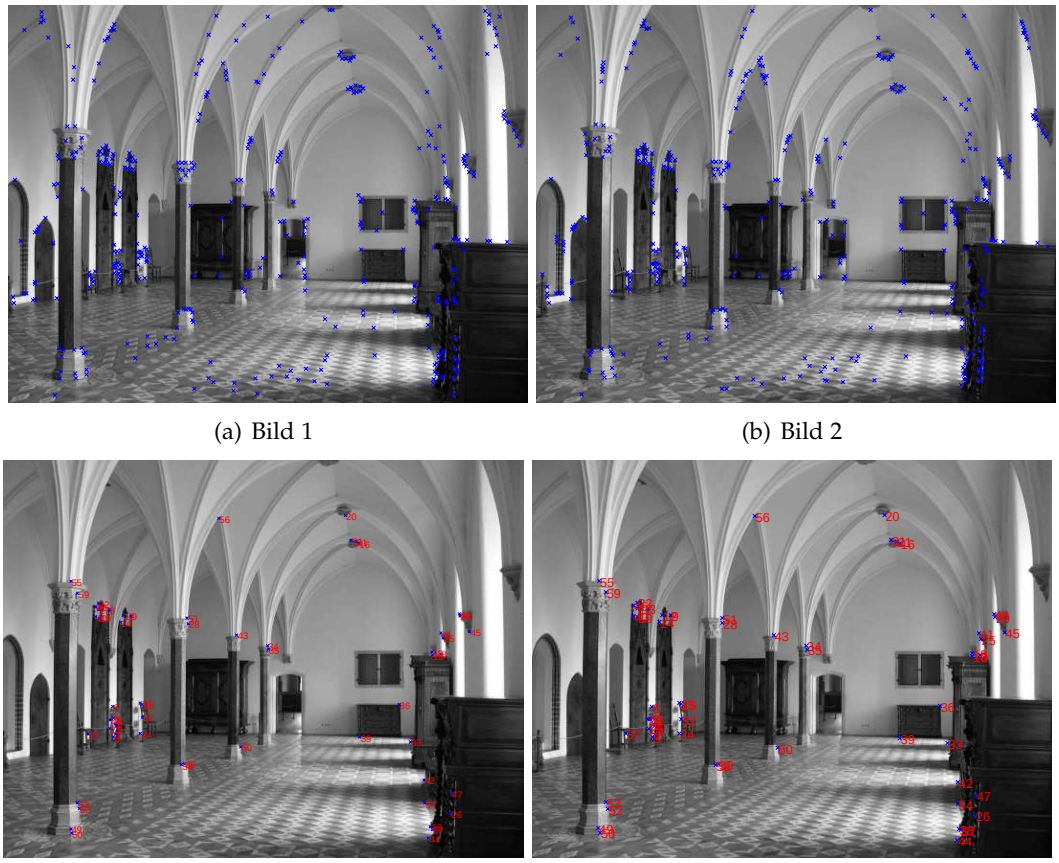
Algorithmus 3: Zusammenfassung der Berechnungsschritte zur Auswahl von Zuordnungen

Punktkorrespondenzen zuerst die Beziehung zwischen den beiden Ansichten, die sich algebraisch durch die Fundamentalmatrix ausdrücken lässt, zu bestimmen. Als Grundlage für die Bestimmung dieser Beziehung dient die *Epipolargleichung* (Gl. 3.1) $\mathbf{m}'^T F \mathbf{m} = 0$.

In Kapitel 3 wurden bereits robuste Schätzverfahren vorgestellt und für die Bestimmung der Epipolargeometrie untersucht. Es wurde gezeigt, dass die Verwendung von ausschließlich linearen Ansätzen wie dem *8-Punkt-Algorithmus* oder dem *normalisierten 8-Punkt-Algorithmus* zu größeren Fehlern führte als die Verwendung von robusten Ansätzen, wenn Ausreißer unter den Punktkorrespondenzen waren. In diesem Abschnitt soll noch einmal der Prozess zur Schätzung der Epipolargeometrie anhand eines der vorgestellten Verfahren zusammengefasst werden.

4.3.1. Initialisierung durch Eliminierung von Ausreißern

Für eine initiale Schätzung der Fundamentalmatrix ist der erste Schritt die Eliminierung von Ausreißern, d.h. schlechte oder falsche Punktkorrespondenzen, die die Schätzergebnisse negativ beeinflussen. Zu diesem Zweck soll der in Abschnitt 3.3.1 vorgestellte RANSAC Algorithmus [48] verwendet werden, der sich sowohl in Applikationen des heutigen



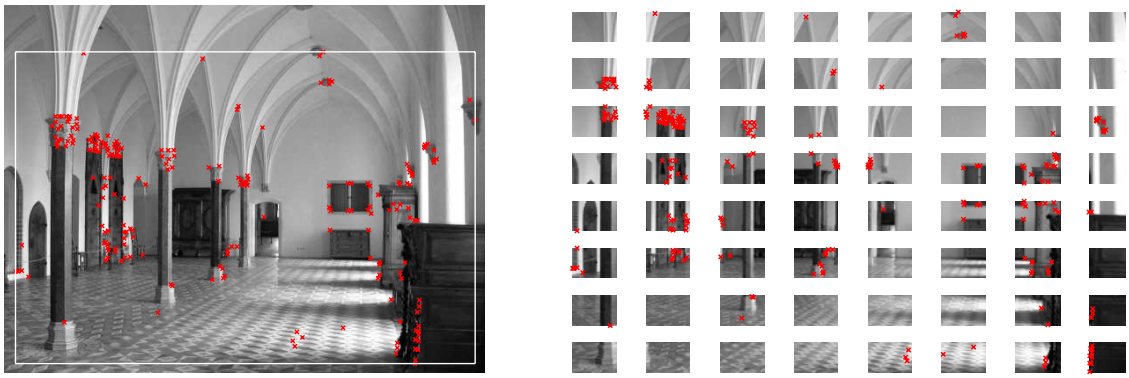
(a) Bild 1 (b) Bild 2

(c) Bild 1 (d) Bild 2

Abbildung 4.8.: Bestimmung eindeutiger Zuordnungen: (a) Extraktion von 400 Merkmalspunkten in Bild 1, (b) Extraktion von 400 Merkmalspunkten in Bild 2, (c) bzw. (d) Zuordnung der Merkmalspunkte in beiden Bildern mit Hilfe der Nachbarschaftsbedingung

Standes der Forschung [2, 24, 60, 130, 142] etabliert hat als auch beim Vergleich robuster Ansätze in Abschnitt 3.4 weitestgehend die besten Ergebnisse lieferte.

Da RANSAC zu den Monte-Carlo-Methoden zählt und zufällig über mehrere Durchläufe acht Punktkorrespondenzen aus dem Datensatz $\mathcal{S} = \{(\mathbf{m}_i \leftrightarrow \mathbf{m}'_i) \,|i = 1, 2, ..., n\}$ auswählt, ist, wie in Abschnitt 4.2.1 beschrieben, die Wahrscheinlichkeit sehr hoch, dass diese innerhalb eines kleinen Bildausschnitts liegen, wenn dort gehäuft Merkmalspunkte aufgrund eines hohen Texturgehalts auftreten. Eine Modifikation des ursprünglichen RANSAC-Algorithmus ist in diesem Fall sinnvoll, da sich die Fundamentalmatrix nicht eindeutig bestimmen lässt, wenn die ausgewählten Punkte z. B. auf einer Ebene liegen. Werden hingegen Punktkorrespondenzen ausgewählt, die mit größerem Abstand voneinander entfernt liegen, sinkt die Wahrscheinlichkeit, dass dieser Fall eintritt. Diese Erweiterung des RANSAC-Algorithmus lässt sich einfach durch Aufteilung des Bildes in gleich große Bildausschnitte erreichen.



(a) Bestimmung eines begrenzenden Fensters um die Merkmalspunkte

(b) Aufteilung des Fensters in 8×8 gleich große Bildausschnitte

Abbildung 4.9.: Modifikation des RANSAC-Algorithmus durch Aufteilung des Bildes in Bereiche

Abbildung 4.9 zeigt ein Beispiel für das linke Bild des Bildpaars *Halle*. Zuerst wird ein rechteckförmiges Fenster um die Merkmalspunkte (z. B. in der linken der beiden Ansichten) gelegt, so dass der linke obere Eckpunkt des Fensters die Koordinaten

$$\tilde{\mathbf{m}}_w^{lo} = \begin{pmatrix} x_w^{lo} \\ y_w^{lo} \end{pmatrix} = \begin{pmatrix} \min\limits_{\forall i}(x_i) \\ \min\limits_{\forall i}(y_i) \end{pmatrix} \tag{4.16}$$

bzw. der rechte untere Eckpunkt die Koordinaten

$$\tilde{\mathbf{m}}_w^{ru} = \begin{pmatrix} x_w^{ru} \\ y_w^{ru} \end{pmatrix} = \begin{pmatrix} \max\limits_{\forall i}(x_i) \\ \max\limits_{\forall i}(y_i) \end{pmatrix} \tag{4.17}$$

besitzt (s. Abbildung 4.9(a)). Anschließend wird das Fenster in gleich große Bereiche aufgeteilt, wie in Abbildung 4.9(b) dargestellt ist. Werden nun zufällig nacheinander acht Merkmalspunkte ausgewählt, wird gleichzeitig überprüft, ob einer der ausgewählten Punkte aus einem Bildausschnitt stammt, in dem bereits ein zuvor ausgewählter Punkt liegt. Ist bereits aus diesem Bildausschnitt ein Punkt zuvor ausgewählt worden, so wird für den aktuellen Merkmalspunkt ein neuer Punkt bestimmt, d.h. alle acht gewählten Merkmalspunkte müssen aus acht unterschiedlichen Bildausschnitten stammen.

In Algorithmus 4 sind die Schritte für eine initiale Schätzung der Fundamentalmatrix $\hat{F}$ und die Eliminierung von Ausreißern zusammengefasst. Die Abbildungen 4.10 und 4.11 zeigen jeweils ein Beispiel für die Eliminierung von Ausreißern. In den oberen beiden Bildern ist jeweils die linke und rechte Ansicht des Bildpaars *Halle* bzw. *Turm* vor der Eliminierung der Ausreißer und in den unteren beiden Bildern das Ergebnis nach der Eliminierung der Ausreißer dargestellt.

Ziel: Eliminierung von Ausreißern und initiale Lösung für die Fundamentalmatrix $\hat{F}$, wobei die gegebene Anzahl n von Punktkorrespondenzen $\{\mathbf{m}_i \leftrightarrow \mathbf{m}'_i\}$ (mit $n > 8$) Ausreißer enthalten kann.

- Festlegung eines Fensters um die Merkmalspunkte in der ersten Ansicht mit der linken oberen Ecke gemäß Gl. 4.16 und der rechten unteren Ecke gemäß Gl. 4.17.
- Aufteilung des Fensters in gleich große Bildausschnitte.
- **Wiederhole für** $k = 1$ **bis** N, wobei N mittels Gl. 3.17 bestimmt wird.
  1. Wähle zufällig einen Satz von 8 Punktkorrespondenzen aus 8 unterschiedlichen Bildausschnitten aus und bestimme die Fundamentalmatrix F_k mit dem normalisierten 8-Punkt-Algorithmus (Algorithmus 2).
  2. Bestimme für jede Punktkorrespondenz den quadratischen Fehler e_i^2 mittels Gl. 3.30 oder Gl. 3.31.
  3. Berechne die robuste Standardabweichung $\hat{\sigma}$ gemäß Gl. 3.20.
  4. Bestimme den Schwellwert T mit $T = 3,84\hat{\sigma}^2$ [68] zur Eliminierung von Ausreißern.
  5. Eliminiere die Ausreißer, deren quadratischer Fehler $e_i^2 > T$ ist.
  6. Optional: Berechne den Anteil an Ausreißern und damit die Anzahl der nötigen Iterationen N gemäß Gl. 3.17 neu.

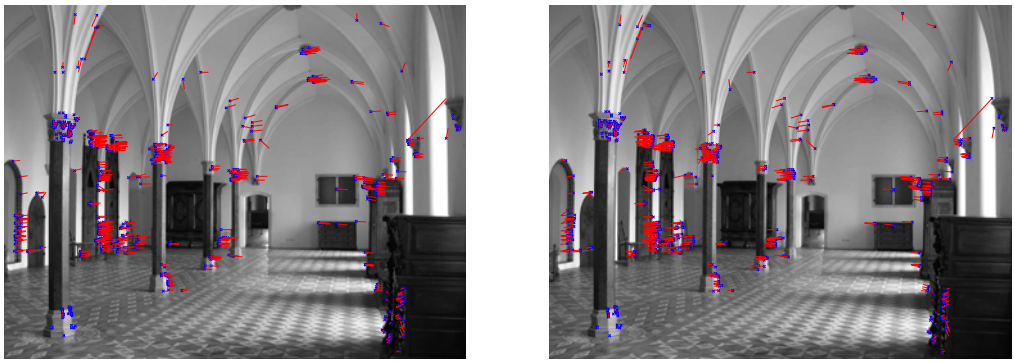
  Ende der Schleife
- Wähle diejenige Fundamentalmatrix F_k, welche die meisten gültigen Punktkorrespondenzen besitzt. Wenn mehr als eine Lösung existiert, wähle diejenige F_k, bei der die Summe der quadratischen Fehler $\sum e_i^2$ der gültigen Punktkorrespondenzen am kleinsten ist.
- Bestimme $\hat{F}$ mit allen gültigen Punktkorrespondenzen über den normalisierten 8-Punkt-Algorithmus (Algorithmus 2).

Ergebnis: Ein Datensatz an gültigen Punktkorrespondenzen und eine initiale Lösung für die Fundamentalmatrix $\hat{F}$.

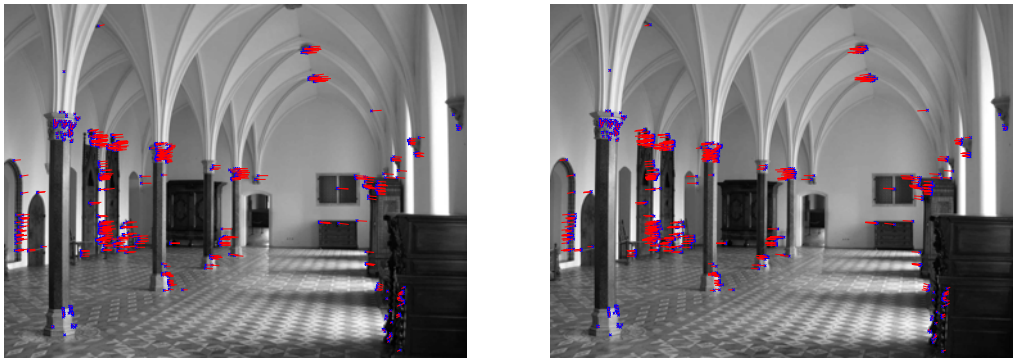
Algorithmus 4: RANSAC-Algorithmus für die robuste Bestimmung der Epipolargeometrie zwischen zwei Ansichten

4.3.2. Optimierung der Schätzergebnisse

Die im vorherigen Abschnitt ermittelten Ergebnisse sollen nun weiter optimiert werden. Als Grundlage für die Optimierung dienen sowohl die initiale Schätzung der Fundamentalmatrix als auch der um die Ausreißer reduzierte Datensatz. Eine Möglichkeit der Optimierung wurde bereits in Kapitel 3 beschrieben, in dem einige nichtlineare Schätzverfahren vorgestellt worden sind. Eine weitere Methode zur Verbesserung der Schätzergebnisse kann über zusätzliche Punktkorrespondenzen erfolgen. Da eine initiale Schätzung der Fundamentalmatrix bereits vorhanden ist, lassen sich neue Merkmalspunkte über die Epipolarbedingung leichter zuordnen, d.h. korrespondierende Punkte müssen auf der entsprechenden Epipolarlinie im anderen Bild liegen. Damit wäre der Suchbereich, der



(a) Bild 1 (b) Bild 2



(c) Bild 1 (d) Bild 2

Abbildung 4.10.: Eliminierung von Ausreißern (Bildpaar *Halle*): (a)-(b) Merkmalspunkte mit Bewegungsvektoren vor der Eliminierung der Ausreißer, (c)-(d) Merkmalspunkte mit Bewegungsvektoren nach der Eliminierung der Ausreißer

zuvor über die Disparitätsbedingung auf ein Suchfenster beschränkt war, nun auf eine Linie bzw. einen kleinen Bereich der Linie begrenzt. Diese Art der Zuordnung von Merkmalspunkten zwischen zwei Ansichten wird auch als *geführte Zuordnung* (engl. *guided matching*) bezeichnet.

Nichtlineare Optimierung

Für die nichtlineare Optimierung der Schätzergebnisse dient vor allem der in Abschnitt 3.2.3 beschriebene Levenberg-Marquardt (LM) Algorithmus. Als zu minimierende Kostenfunktion wird vorrangig entweder der Sampson-Fehler nach Gl. 3.30 oder der 2D-Reprojektionsfehler nach Gl. 3.31 verwendet. Eine detaillierte Beschreibung des LM-Algorithmus ist im Anhang A.2 zu finden.

In Tabelle 4.1 sind für drei verschiedene Bildpaare die mittleren Abstände der Merkmalspunkte zu den Epipolarlinien nach der Schätzung der Fundamentalmatrix angegeben. Dabei werden die Ergebnisse der verschiedenen Prozess-Schritte verglichen und gegenüber gestellt. Der erste Schritt ist die Eliminierung von Ausreißern und die Bestimmung

(a) Bild 1 (b) Bild 2

(c) Bild 1 (d) Bild 2

Abbildung 4.11.: Eliminierung von Ausreißern (Bildpaar *Turm*): (a)-(b) Merkmalspunkte mit Bewegungsvektoren vor der Eliminierung der Ausreißer, (c)-(d) Merkmalspunkte mit Bewegungsvektoren nach der Eliminierung der Ausreißer

Tabelle 4.1.: Robuste Schätzung der Fundamentalmatrix ohne und mit nichtlinearer Minimierung. Als Minimierungskriterium wurde der Sampson-Fehler verwendet. Nach der Eliminierung der Ausreißer blieben beim Bildpaar *Korridor* (Auflösung: 512×512 Pixel) 497 von ursprünglich 518 , *Halle* (Auflösung: 2185×1672 Pixel) 610 von 669 und *Turm* (Auflösung: 1200×2020 Pixel) 1040 von 1188 gültige Punktkorrespondenzen übrig.

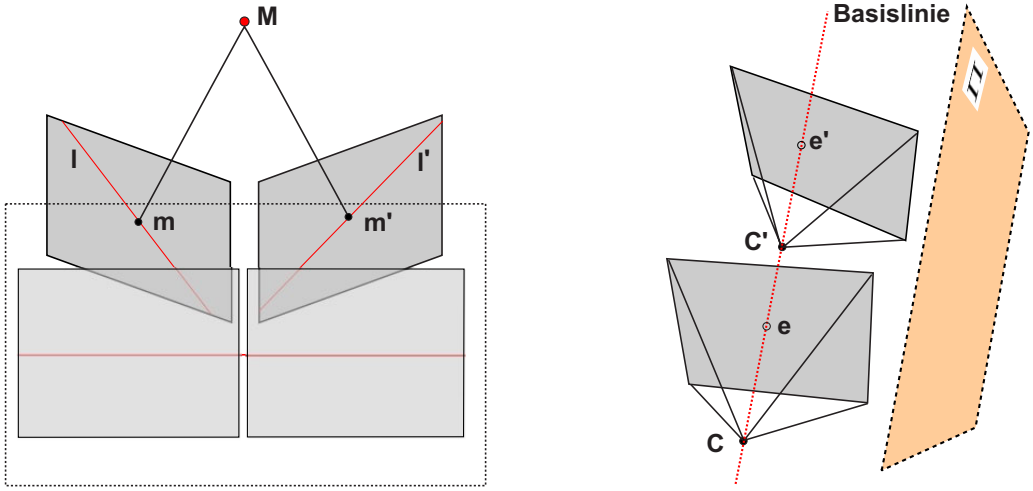
	Mittlerer Abstand der Punkte zu den Epipolarlinien in Pixeln		
Bildpaar	*Korridor*	*Halle*	*Turm*
RANSAC	1,0515	1,2478	1,0237
RANSAC + n8pt	0,5353	0,6559	0,6469
RANSAC + n8pt + LM	0,4623	0,6541	0,5892

der initialen Fundamentalmatrix F_k mittels RANSAC nach Algorithmus 4. Im zweiten Schritt wird mit den gültigen Punktkorrespondenzen (nach Eliminierung der Ausreißer) über den normalisierten 8-Punkt Algorithmus die Fundamentalmatrix $\hat{F}$ neu berechnet. Der letzte Schritt beinhaltet die nichtlineare Optimierung der Parameter der Fundamentalmatrix F_{LM} mit Hilfe des LM-Algorithmus. Als Kostenfunktion wurde der Sampson-Fehler verwendet. Für alle gültigen Punktkorrespondenzen wurde der mittlere Abstand zu den entsprechenden Epipolarlinien über die drei Fundamentralmatrizen F_k, $\hat{F}$ und F_{LM} jeweils berechnet.

4.3.3. Rektifizierung

Durch zusätzliche Punktkorrespondenzen in beiden Ansichten können die Schätzergebnisse ebenfalls weiter optimiert werden. Eine Möglichkeit den Suchbereich und damit die Mehrdeutigkeit von Punktkorrespondenzen zu reduzieren, ist die Ausnutzung der Epipolarbedingung. Basierend auf der Idee, den Suchbereich auf eine Linie zu reduzieren, wurden Verfahren entwickelt, die unter dem Namen *Rektifizierung* (engl. *rectification*) bekannt sind [6,113,133,134,143,154]. Bei der Rektifizierung werden die Bilder so transformiert, dass die Epipolarlinien horizontal angeordnet sind. Dies hat vor allem den Vorteil, dass bei der Suche nach eindeutigen Zuordnungen nicht jeweils die Geradengleichung der Epipolarlinie mitberücksichtigt werden muss, da durch die Rektifizierung Punktkorrespondenzen in der gleichen Zeile der beiden Bilder liegen (siehe Abbildung 4.12(a)).

Bei standard Rektifizierungsverfahren, die auch als *planare Rektifizierung* bezeichnet werden, wird für beide Ansichten jeweils eine Homographie berechnet, mit der die Bilder auf eine gemeinsame Ebene transformiert werden können [6, 113, 134]. Geometrisch



(a) Prinzip der Rektifizierung eines Bildpaares

(b) Probleme der planaren Rektifizierung bei vorwärts gerichteter Kamerabewegung

Abbildung 4.12.: Planare Rektifizierung

werden bei dieser Transformation die Epipole ins Unendliche verschoben, so dass die Epipolarlinien parallel zueinander sind. Ein großer Vorteil dieser Verfahren liegt darin, dass in vielen Anwendungsszenarien (z. B. [5, 159]), bei denen sich die Aufnahmeeinrichtung der Kameras nicht ändert, die Transformationsparameter nur einmal berechnet werden müssen. Allerdings sind diese Verfahren nicht allgemein nutzbar bzw. nicht immer einsetzbar. Abbildung 4.12(b) zeigt ein Beispiel für die Anordnung zweier Kameras, bei der eine planare Rektifizierung nicht möglich wäre, da die rektifizierten Bilder unendlich groß werden würden, wenn bei der Transformation die Epipole, die hier innerhalb der Bilder liegen, ins Unendliche verschoben werden. Auch für den Fall, dass die Epipole nicht innerhalb der Bilder sondern sehr nahe am Bildrand liegen, würden die rektifizierten Bilder zumindest sehr groß werden. Dieses Problem wird z. B. in [154] umgangen, indem die Bilder nicht auf eine Ebene sondern auf einen Zylinder transformiert werden.

Ein anderer Ansatz der Rektifizierung wird in [143] und [133] vorgeschlagen, der auch als *polare Rektifizierung* bezeichnet wird. Dieser Ansatz benötigt lediglich die Fundamentalmatrix und ist allgemein für jede Art der Kamerabewegung einsetzbar, z. B. bei Videosequenzen, bei denen sich die Kamera zumindest zeitweise vor- oder rückwärts gerichtet bewegt hat, oder bei Fotoaufnahmen, bei der die Aufnahmepositionen hintereinander angeordnet sind. Bei der polaren Rektifizierung werden die Bilder um den Epipol mithilfe polarer Koordinaten reparametrisiert.

Der erste Schritt der polaren Rektifizierung ist die Bestimmung der beiden extremen Epipolarlinien (diejenigen Epipolarlinien, die die äußersten Bildecken schneiden) für jedes Bild, wie in Abbildung 4.13 dargestellt wird (l_1 und l_2 im linken und l'_3 und l'_4 im rechten Bild). Die gemeinsame Region, d.h. der Bildbereich, der in beiden Ansichten sichtbar

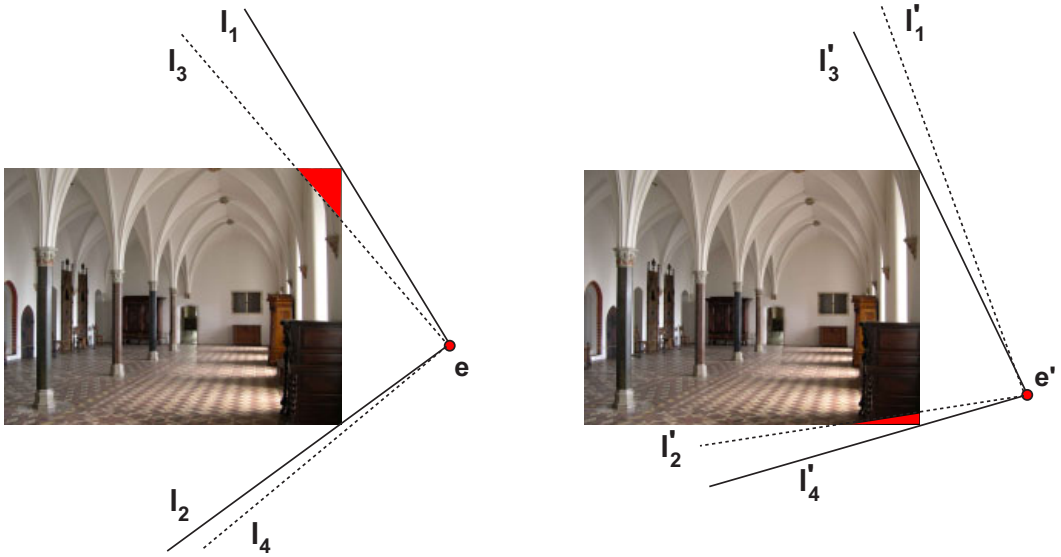
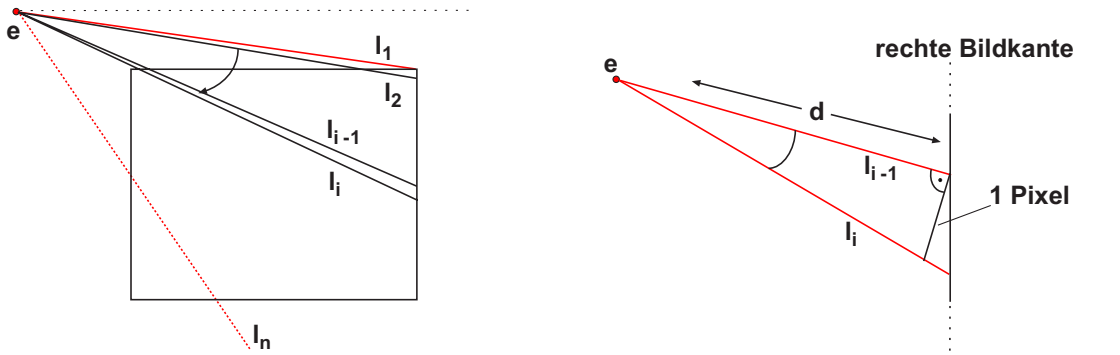


Abbildung 4.13.: Polare Rektifizierung: Bereiche, die in beiden Bildern sichtbar sind



(a) Rektifizierung in Winkelschritten

(b) Minimal erforderlicher Winkel zwischen zwei aufeinander folgenden Epipolarlinien

Abbildung 4.14.: Prinzip der polaren Rektifizierung

ist, wird durch Transformation der extremen Epipolarlinien in das entsprechend andere Bild ermittelt (siehe Abbildung 4.13). Jede Zeile der polar rektifizierten Bilder erhält man nun durch Scannen der Epipolarlinien mit entsprechenden Winkelschritten innerhalb der gemeinsamen Region, d.h. die Epipolarlinien, die in beiden Bildern innerhalb des sichtbaren Bildbereichs liegen, werden zeilenweise untereinander abgelegt. Abbildung 4.14(a) zeigt dafür ein Beispiel. Um dabei Pixel-Kompressionen zu vermeiden, muss zwischen den zu scannenden Epipolarlinien der minimal erforderliche Winkel bestimmt werden. Der kritische Bereich liegt dabei an der am weitesten entfernten Bildkante, die von der Epipolarlinie geschnitten wird (siehe Abbildung 4.14(b)). Der Abstand zwischen der zu scannenden Epipolarlinie und der zuvor gescannten Epipolarlinie sollte an dieser Bildkante genau ein Pixel betragen, so dass der Winkel berechnet werden kann. Im Gegensatz zur

planaren Rektifizierung steht die maximale Größe eines rektifizierten Bildes bereits fest. Die maximale vertikale Auflösung (maximale Anzahl an Zeilen) erhält man, wenn der Epipol innerhalb des Bildes liegen würde, da die Epipolarlinien mit 360° um den Epipol gescannt werden müssten:

$$x_{max} = 2 \cdot (y_{orig} + x_{orig}), \tag{4.18}$$

mit y_{orig} und x_{orig} die horizontale und vertikale Auflösung des Originalbildes. Die maximale horizontale Auflösung ergibt sich, wenn der Epipol außerhalb des Bildes auf der Bilddiagonalen liegen würde:

$$y_{max} = \sqrt{y_{orig}^2 + x_{orig}^2} \tag{4.19}$$

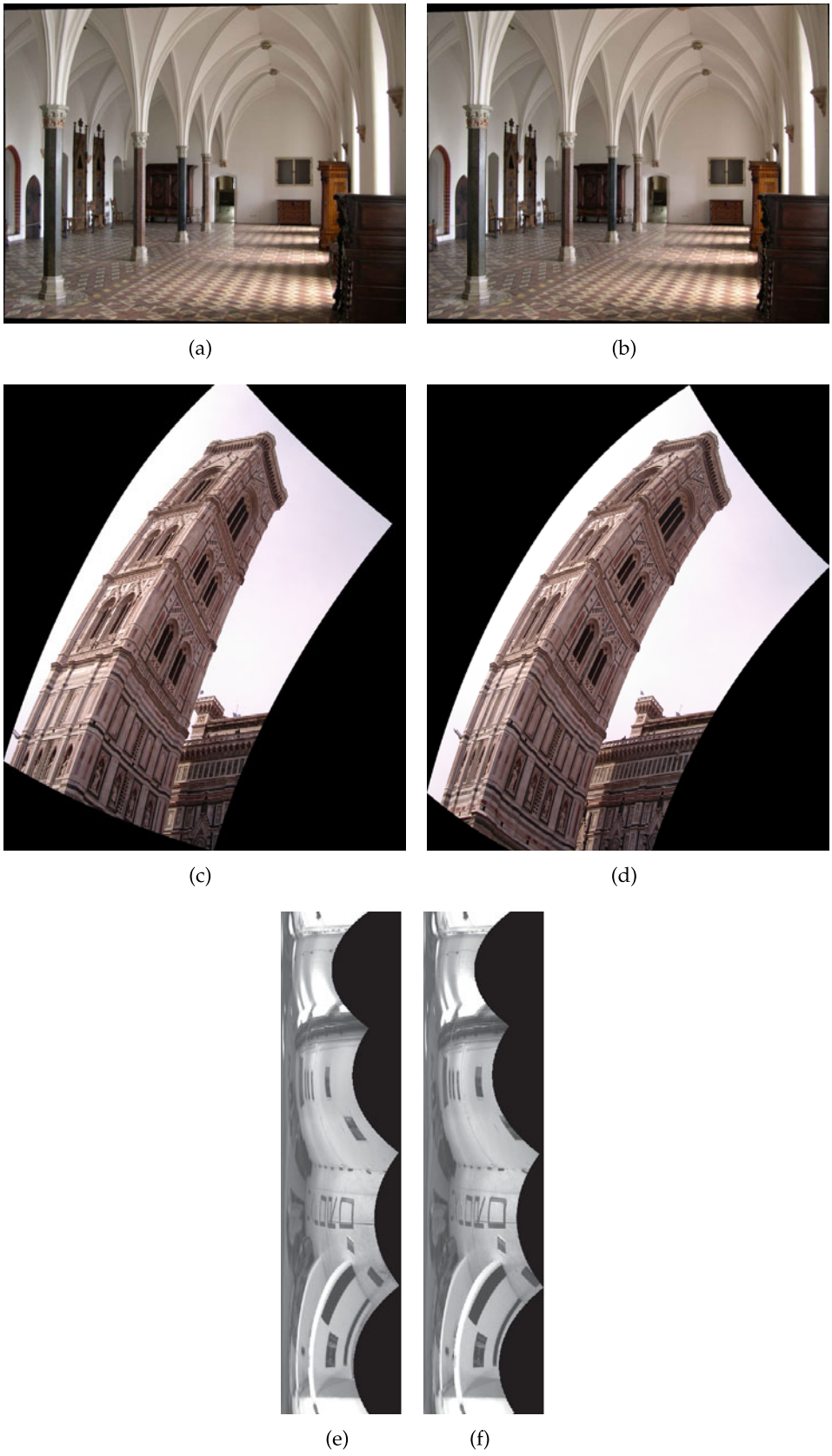
Beispiel: Ein Bildpaar habe jeweils die Auflösung von 720×576. Je nach Lage des Epipols wäre die maximale vertikale Auflösung $x_{max} = 2 \cdot (720 + 576) = 2592$ Pixel (Epipol liegt innerhalb des Bildes) und die maximale horizontale Auflösung $y_{max} = \sqrt{720^2 + 576^2} = 922$ Pixel (Epipol liegt außerhalb des Bildes).

In Abbildung 4.15 sind die Ergebnisse der polaren Rektifizierung dreier Bildpaare dargestellt. Die Originalauflösung für das Bildpaar *Halle* ist 2185×1672 Pixel, für *Turm* 1200×2020 Pixel und für *Korridor* 512×512 Pixel. Die Kameraausrichtung bei dem Bildpaar *Halle* war annähernd parallel, d.h. die Epipole liegen weit außerhalb der beiden Bilder. Beim Bildpaar *Korridor* liegt der Epipol jeweils innerhalb der Bilder.

4.4. Selbstkalibrierung

Da die internen Kameraparameter in der Regel nicht bekannt sind, können diese über Selbstkalibrierungsverfahren geschätzt werden. Die erste und wohl bekannteste Kalibrierungsmethode führt über die Lösung der Kruppa-Gleichungen und wurde von Faugeras et al. [40] entwickelt. In [118] wurde dieser Ansatz durch ein nichtlineares Optimierungsverfahren erweitert. Eine weitere Kalibrierungstechnik wurde von Pollefeys und van Gool [145] entwickelt. Sie verwendeten einen stratifizierten Ansatz, bei dem die Kameraparameter durch Erweiterung einer projektiven Rekonstruktion über eine affine bis zur Euklidischen Rekonstruktion geschätzt wurden. Diese Methode gilt als sehr viel stabiler und robuster gegenüber kritischen Kamerabewegungen. Allerdings werden auch hier nichtlineare Optimierungsverfahren für die affine Rekonstruktion benötigt.

In vielen Fällen ist die fokale Länge die einzige Unbekannte der Kamerakalibreirungsmatrix. Unter dieser Voraussetzung können lineare Methoden verwendet werden, wie sie u.a. in [170] vorgeschlagen werden. Eine einfache und sehr effektive Kalibrierungstechnik ist in [65] und [126] zu finden, bei der Bedingungen an die Singulärwerte der Essentialmatrizen gestellt werden.



(a) (b)

(c) (d)

(e) (f)

Abbildung 4.15.: Polar rektifizierte Bildpaare: (a)-(b) Bildpaar *Halle* (Auflösung: 2195 × 1742), (c)-(d) Bildpaar *Turm* (Auflösung: 2163 × 2506) und (e)-(f) Bildpaar *Korridor* (Auflösung: 426 × 1793)

4.5. Bestimmung der Rotation und Translation

Nachdem die Fundamentalmatrix zwischen zwei Ansichten geschätzt wurde, geht es nun darum, die Kameraprojektionsmatrizen P und P' zu bestimmen, so dass die 3D-Raumpunkte $\mathbf{M}$ entsprechend Gl. 4.14 in den beiden Ansichten über

$$\mathbf{m} = P\mathbf{M} = K\left[\underline{I}|\mathbf{0}\right]\mathbf{M} \tag{4.20}$$

$$\mathbf{m}' = P'\mathbf{M} = K'\left[R|-R\tilde{\mathbf{C}}\right]\mathbf{M} \tag{4.21}$$

abgebildet werden. K bzw. K' sind die Kamerakalibrierungsmatrizen mit den intrinsischen Kameraparametern, $\underline{I}$ ist eine 3×3 Einheitsmatrix, d.h. das Referenzkoordinatensystem (bzw. Weltkoordinatensystem) ist hier in die erste Kamera gelegt, R ist die Rotationsmatrix und $-R\tilde{\mathbf{C}} = \mathbf{t}$ ist der Translationsvektor, der die relative Position der zweiten Kamera bezüglich der ersten Kamera angibt, wobei $\tilde{\mathbf{C}}$ das Kamerazentrum der zweiten Kamera in inhomogenen Koordinaten ist.

Unter der Annahme, dass die intrinsischen Kameraparameter und damit die Kamerakalibrierungsmatrizen bekannt sind, soll nun die relative Position und Orientierung der Kameras bestimmt werden. Die normalisierten Bildkoordinaten der Punkte $\mathbf{m}$ und $\mathbf{m}'$ lassen sich mit

$$\hat{\mathbf{m}} = K^{-1}\mathbf{m}$$

$$\hat{\mathbf{m}}' = K'^{-1}\mathbf{m}'$$

angeben. Die Epipolargleichung kann damit zu

$$\mathbf{m}'^{T}F\mathbf{m} = \hat{\mathbf{m}}'^{T}K'^{T}FK\hat{\mathbf{m}} = \hat{\mathbf{m}}'^{T}E\hat{\mathbf{m}} = 0 \tag{4.22}$$

umgeformt werden, wobei E die *Essentialmatrix* mit

$$E = K'^{T}FK \tag{4.23}$$

ist. Die Essentialmatrix kann mit Gl. 4.23 direkt aus der Fundamentalmatrix und den Kamerakalibrierungsmatrizen berechnet werden und hat die Form

$$E = [t]_x R = SR \tag{4.24}$$

[112]. Die Matrix $S = [t]_x$ ist eine *antisymmetrische* Matrix mit

$$S = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}. \tag{4.25}$$

Zur Berechnung der zweiten Kameramatrix P' können die antisymmetrische Matrix und die Rotationsmatrix durch Faktorisierung bestimmt werden. Hartley schlägt in [65] eine numerisch stabile Methode vor, die auf der Singulärwertzerlegung beruht.

4.6. Triangulation

Ein wesentlicher Prozess der 3D-Rekonstruktion ist die Triangulation, welche die 3D-Koordinaten der Punktkorrespondenzen in den Bildern bestimmt. Physikalisch ist die Triangulation nichts anderes als die Anwendung der Strahlensätze. Sind die Kameraprojektionsmatrizen und die 2D-Koordinaten korrespondierender Bildpunkte bekannt, dann lässt sich der 3D-Raumpunkt als Schnittpunkt der rückprojezierten Strahlen von den Kamerazentren durch die Bildpunkte bestimmen:

$$\mathbf{M} = \mathcal{T}(\mathbf{m}, \mathbf{m}', P, P'), \tag{4.26}$$

wobei $\mathcal{T}$ für das verwendete Triangulationsverfahren steht. Da aber die 2D-Koordinaten der Bildpunkte durch Rauscheinflüsse fehlerbehaftet sind, treffen sich diese Strahlen nicht in einem Punkt im dreidimensionalen Raum. Lineare Triangulationsverfahren [38, 68] lösen dieses Problem, ähnlich dem in Abschnitt 3.1.1 vorgestellten 8-Punkt-Algorithmus, mittels Singulärwertzerlegung. In den folgenden beiden Abschnitten werden zwei weitere gängige Verfahren zur Triangulation kurz vorgestellt: die Mittelpunkt-Methode [10] und *polynomiale* Triangulation [67], die auch unter dem Namen *optimale* Triangulation bekannt ist [68].

4.6.1. Mittelpunkt-Methode

Ein sehr populäres und anschauliches Triangulationsverfahren stellt die Mittelpunkt-Methode [10] dar, bei der die Strahlen durch die Kamerazentren und die Bildpunkte berechnet und der Mittelpunkt an der Stelle des kleinsten Abstands der Strahlen im dreidimensionalen Raum bestimmt wird. Abbildung 4.16 zeigt ein Beispiel für die Rückprojektion der Strahlen durch die fehlerbehafteten Punktkorrespondenzen (weiße Punkte), den berechneten Mittelpunkt und die Projektion des Mittelpunktes in die Bildebenen (rote Punkte). Die Rückprojektion der Strahlen kann nach Gl. 4.20 über die Pseudoinversen der Kameraprojektionsmatrizen P^+ und P'^+ mit

$$P^+ = P^T(PP^T)^{-1} \tag{4.27}$$

ermittelt werden. Für den Bildpunkt $\mathbf{m}$ ergibt sich somit der Strahl

$$\mathbf{M}(\alpha) = P^+\mathbf{m} + \alpha\mathbf{C}, \tag{4.28}$$

auf dem der 3D-Raumpunkt $\mathbf{M}$ liegt. α ist dabei der unbekannte Skalierungsfaktor. Entsprechend lässt sich der Strahl für den Bildpunkt $\mathbf{m}'$ erzeugen. Der 3D-Raumpunkt ist dann der Mittelpunkt der Strecke im 3D-Raum, die senkrecht auf den beiden Strahlen steht. Diese Methode ist allerdings nur für eine Euklidische Rekonstruktion geeignet, da der Mittelpunkt durch das Fällen eines Lots auf beide Strahlen ermittelt wird. Rechtwinkligkeit ist allerdings keine Eigenschaft der affinen und projektiven Geometrie.

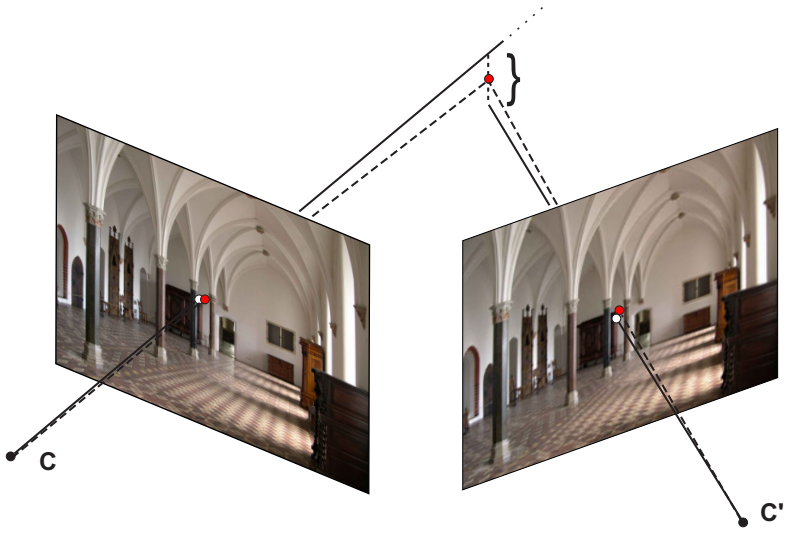


Abbildung 4.16.: Triangulation mit der Mittelpunkt-Methode

4.6.2. Optimale Triangulation

Ein projektiv invariantes Triangulationsverfahren stellt die polynomiale Triangulation [67] dar. Ziel der polynomialen Triangulation ist es, anstelle der korrespondierenden Punkte $\{\hat{\mathbf{m}} \leftrightarrow \hat{\mathbf{m}}'\}$ Punktkorrespondenzen $\{\mathbf{m}_p \leftrightarrow \mathbf{m}'_p\}$ zu finden, die die Epipolargleichung erfüllen und die Kostenfunktion

$$E(\mathbf{m}_p, \mathbf{m}'_p) = d(\hat{\mathbf{m}}, \mathbf{m}_p)^2 + d(\hat{\mathbf{m}}', \mathbf{m}'_p)^2 \tag{4.29}$$

minimieren, so dass sich die rückprojezierten Strahlen im dreidimensionalen Raum schneiden. Die Minimierung erfolgt somit in der Bildebene, wodurch die projektive Invarianz gewährleistet wird. Abbildung 4.17 zeigt ein Beispiel für die polynomiale Triangulation. Die gestrichelten Linien stellen die Strahlen des exakten 3D-Raumpunktes und die durchgezogenen Linien die Strahlen des geschätzten 3D-Punktes dar.

Da sich die rückprojezierten Strahlen nur schneiden, wenn die korrespondierenden Punkte direkt auf den Epipolarlinien liegen, kann das Minimierungsproblem auch mit

$$E(t) = d\left(\hat{\mathbf{m}}, \mathbf{l}(t)\right)^2 + d\left(\hat{\mathbf{m}}', \mathbf{l}'(t)\right)^2 \tag{4.30}$$

angegeben werden, d.h. es wird über den Parameter t nach denjenigen Epipolarlinien $\mathbf{l}$ und $\mathbf{l}'$ gesucht, die Gl. 4.30 minimieren. Hartley und Sturm zeigen in [67], dass $d\left(\hat{\mathbf{m}}, \mathbf{l}(t)\right)^2$ bzw. $d\left(\hat{\mathbf{m}}', \mathbf{l}'(t)\right)^2$ von t abhängige gebrochenrationale Funktionen sind und dass das Minimierungsproblem durch Berechnung reeller Nullstellen eines Polynoms 6. Ordnung gelöst werden kann.

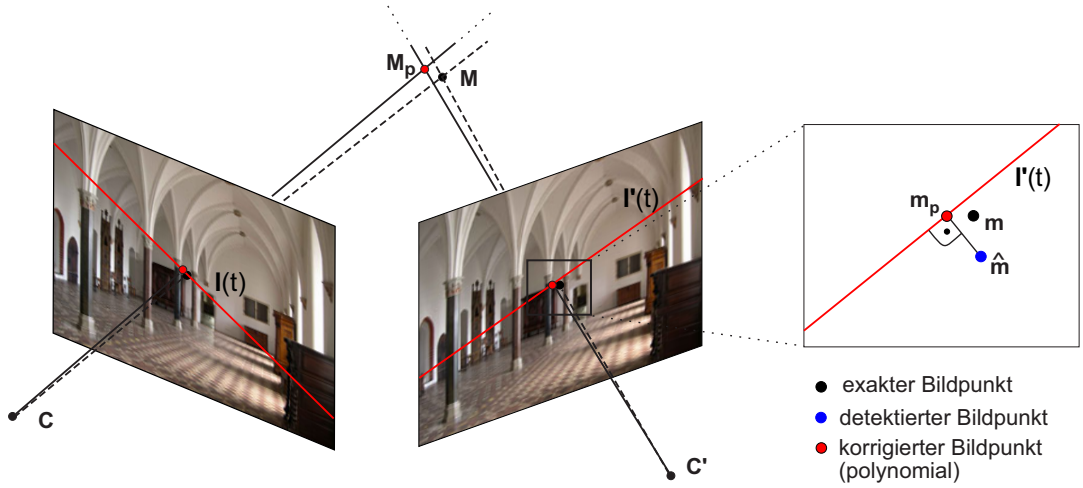


Abbildung 4.17.: Optimale Triangulation

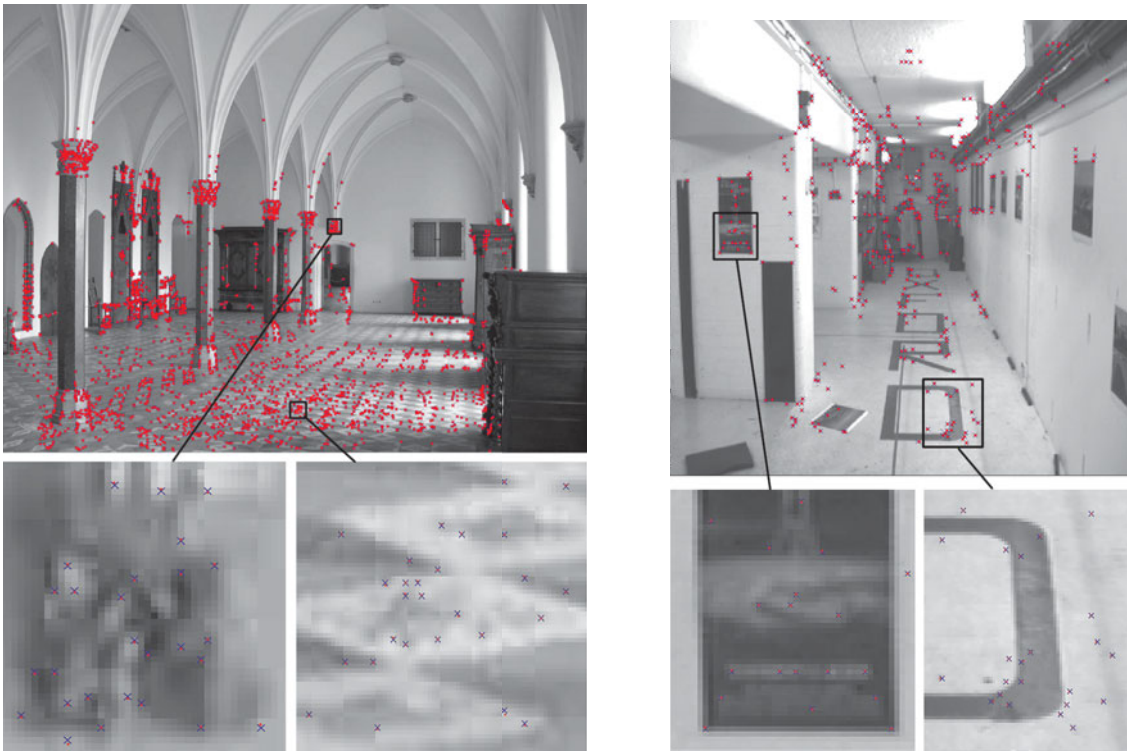
4.7. Ergebnisse

Die beschriebenen Verfahren zur Rekonstruktion der 3D-Struktur und der Kameraprojektionsmatrizen wurden für fünf Bildpaare angewendet. In Tabelle 4.2 sind die mittleren Reprojektionsfehler in Pixeln, die Anzahl der rekonstruierten 3D-Punkte und die jeweilige Auflösung der Bilder aufgeführt. Die Anzahl der rekonstruierten 3D-Punkte hängt im großen Maße von der Bildauflösung ab. Zudem spielt der Texturgehalt eine große Rolle. Das Bildpaar *Halle* enthält z. B. viele homogene Bildbereiche. Bedingt durch starke Beleuchtungsunterschiede und durch eine leichte Rotation der Kameras um die Hauptachsen ist die Anzahl der rekonstruierten 3D-Punkte für das Bildpaar *Turm* relativ gering.

Tabelle 4.2.: Mittlere Reprojektionsfehler für unterschiedliche Bildpaare

Bildpaar	Anzahl rekonstruierter 3D-Punkte	Reprojektionsfehler in Pixeln	Bildauflösung in Pixeln
Halle	3422	0,36	2185×1672
Turm	1029	0,33	1200×2020
Korridor	509	0,26	512×512
Rathaus	2928	0,22	800×533
Totem	1684	0,31	768×512

In Abbildung 4.18 sind die Reprojektionsfehler der rekonstruierten 3D-Punkte visuell für die Bildpaare *Halle* und *Korridor* dargestellt. Die Abbildung zeigt jeweils eine Ansicht der Bildpaare mit den detektierten Merkmalspunkten (blaue Kreuze) und den reproje-



(a) Bildpaar *Halle* (b) Bildpaar *Korridor*

Abbildung 4.18.: Darstellung der Reprojektionsfehler (rot: reprojezierter 3D-Punkt, blau: detektierter Merkmalspunkt)

zierten 3D-Punkten (rote Punkte) und zwei vergrößerte Ausschnitte aus diesen Bildern.

Für die drei weiteren Bildpaare (*Turm*, *Rathaus* und *Totem*) sind die 3D-Rekonstuktionsergebnisse in den Abbildungen 4.19 bis 4.21 dargestellt. Auf der linken Seite der Abbildungen werden jeweils die verwendeten Bilder gezeigt (für das Bildpaar *Turm* wurde nur die Ansicht der linken Kamera dargestellt) und auf der rechten Seite entsprechend die 3D-Rekonstruktion der Szenen aus zwei unterschiedlichen Perspektiven mit den Kamerapositionen und -orientierungen. Rechts unten befindet sich jeweils eine Draufsicht auf die Rekonstruktion aus der Vogelperspektive.

4.8. Kapitelzusammenfassung

In diesem Kapitel wurden verschiedene aktuelle Ansätze und Verfahren der 3D-Rekonstruktion aus zwei Ansichten vorgestellt. Die Rekonstruktion der aufgenommenen Szene und die Bestimmung der Kamerapositionen und -orientierungen basierte auf der Detektion von Merkmalsobjekten und deren Zuordnung. Als Merkmalsobjekte wurden Punkte gewählt, die einen hohen Gradienten innerhalb eines Fensters aufwiesen, da sich

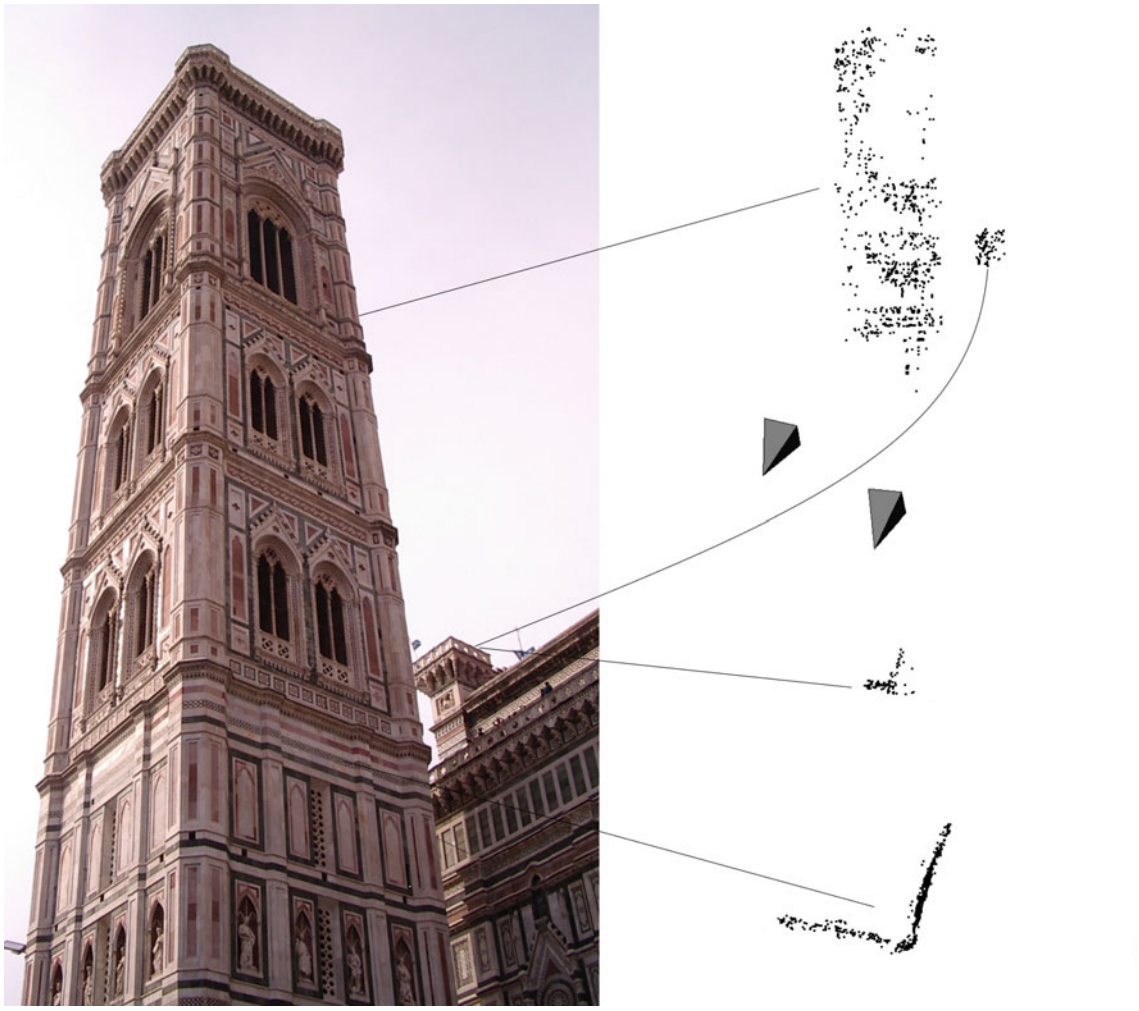


Abbildung 4.19.: Rekonstruktionsergebnis für das Bildpaar *Turm*

diese gerade im Hinblick auf das nächste Kapitel speziell in Videosequenzen leicht zuordnen bzw. verfolgen lassen. Die Zuordnung der Merkmalspunkte erfolgte über ein Verfahren, welches auf der Einzigartigkeits-, Disparitäts- und Nachbarschaftsbedingung basiert.

Ferner wurde ein robustes Verfahren zur Schätzung der Epipolargeometrie vorgestellt, dass zur Initialisierung und zur Eliminierung von Ausreißern RANSAC verwendet und anschließend eine nichtlineare Optimierung über den Sampson-Fehler vornimmt. Um weitere Punktkorrespondenzen zu generieren bzw. um die Schätzergebnisse weiter zu verbessern, konnte zusätzlich die Epipolarbedingung eingesetzt werden, welche den Suchbereich auf die Epipolarlinien beschränkt. Durch polare Rektifizierung der Bilder wurden die Epipolarlinien horizontal angeordnet, so dass sich der Suchbereich auf die entsprechenden Zeilen reduzierte.

Für die Bestimmung der Rotation und Translation der Kameras wurde vorausgesetzt, dass die intrinsischen Kameraparameter durch Selbstkalibrierungsverfahren ermittelt wurden. Eine weitere Möglichkeit, die interenen Kameraparameter zu bestimmen, die

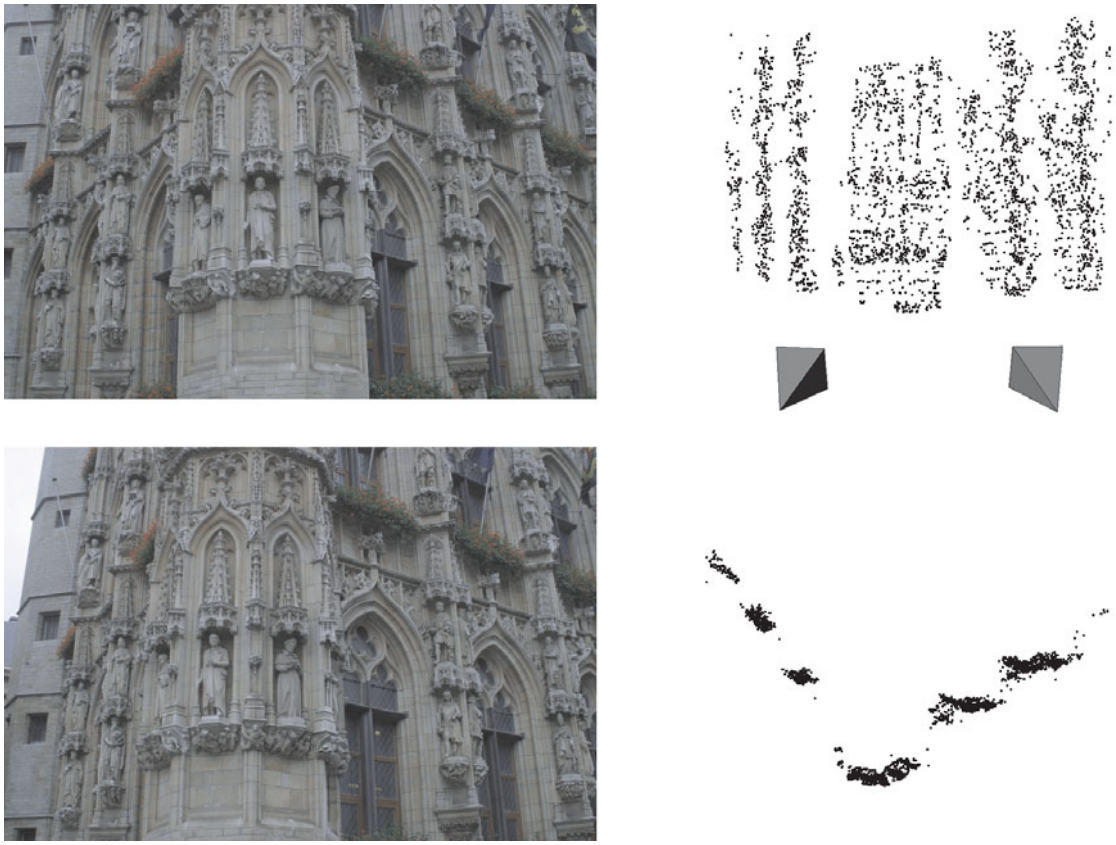


Abbildung 4.20.: Rekonstruktionsergebnis für das Bildpaar *Rathaus*

hier aber nicht weiter beschrieben wurde, ist die manuelle Kamerakalibrierung. Manuelle Kalibrierungstechniken werden in der Regel kurz vor oder nach den jeweiligen Aufnahmen angewendet und benötigen zusätzliche Hilfsmittel wie Kalibrierungsmuster.

Die Rekonstruktion der Szene bzw. die Bestimmung der 3D-Koordinaten der Merkmalspunkte erfolgte über die polynomiale Triangulation, welche auch als optimale Triangulation bezeichnet wird. Das zu lösende Minimierungsproblem wurde dabei nicht im dreidimensionalen Raum sondern durch Parametrisierung der Epipolarlinien in den Bildebenen durchgeführt.

Abschließend wurden die Rekonstruktionsergebnisse von fünf verwendeten Bildpaaren anhand der Reprojektionsfehler und visuell durch Aufnahmen der entsprechenden 3D-Modelle aus unterschiedlichen Perspektiven dargelegt.

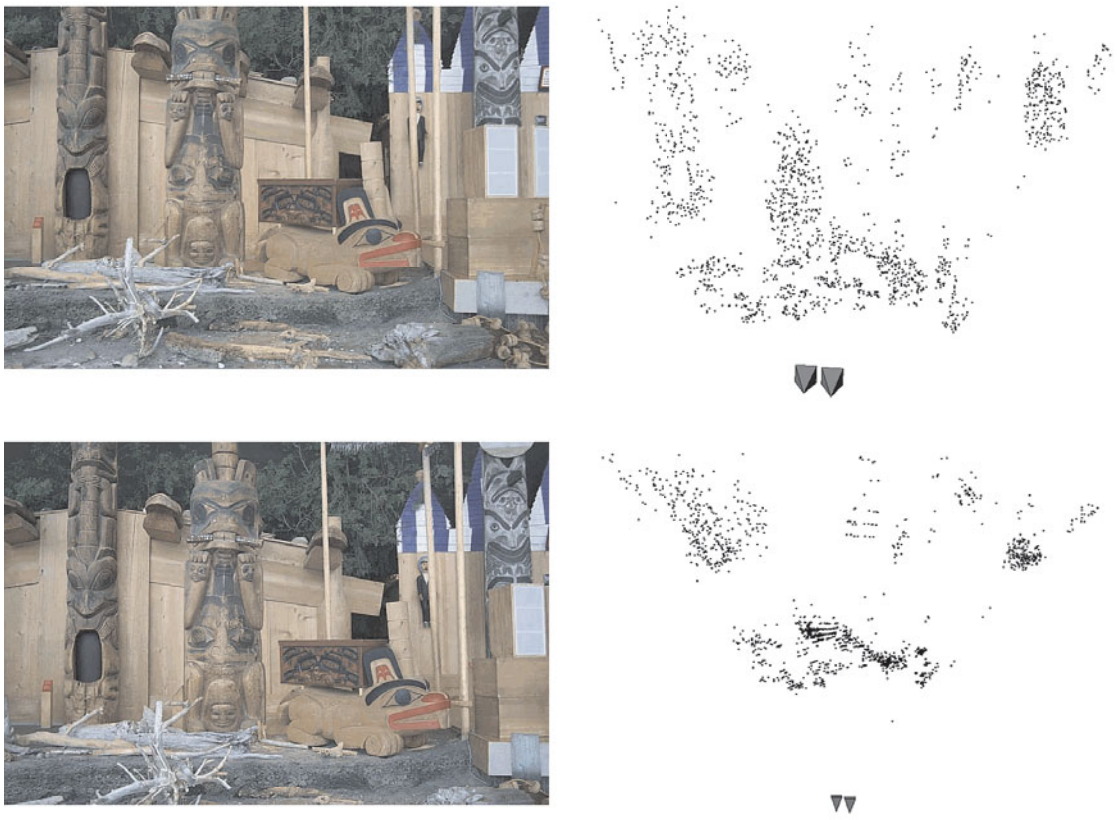


Abbildung 4.21.: Rekonstruktionsergebnis für das Bildpaar *Totem*

5. 3D-Rekonstruktion aus mehreren Ansichten

In dem vorangegangenen Kapitel wurde die Rekonstruktion einer dreidimensionalen Szene mithilfe von zwei Kameraansichten, die z. B. mit einer Fotokamera aufgenommen wurden, beschrieben. Dabei wurde speziell auf Merkmalspunkte, die innerhalb eines kleinen Fensters einen hohen Texturgehalt aufweisen, eingegangen. In diesem Abschnitt wird die Rekonstruktion einer dreidimensionalen Szene mithilfe mehrerer Ansichten vorgestellt. Als Grundlage sollen dabei Ansichten verwendet werden, die durch die Aufnahme mit einer Videokamera entstanden sind. Aufeinander folgende Bilder einer Videosequenz besitzen allgemein große überlappende Bereiche und damit eine große Ähnlichkeit untereinander.

5.1. Problemstellung und Überblick über die Methode

Die Tatsache, dass aufeinander folgende Bilder einer Videosequenz einen hohen Grad an Korrelation aufweisen, lässt sich durch eine langsame Änderung der Szene bzw. durch die Bewegung der Kamera begründen. In der Regel beträgt die Bildrate bei Videofilmen 25 bzw. 30 Bilder pro Sekunde (PAL[1] bzw. NTSC[2]). Demnach ist der zeitliche Abstand zwischen zwei aufeinander folgenden Bildern einer Sequenz $\frac{1}{25}$ bzw. $\frac{1}{30}$ Sekunde: eine Zeitspanne, in der bei realen Aufnahmen tatsächlich nur geringe Änderungen des Bildinhalts auftreten.

In Anlehnung an die im vorangegangenen Abschnitt erklärten Probleme bei der Rekonstruktion aus zwei Ansichten kann man sowohl eine positive als auch eine negative Eigenschaft von Videosequenzen entdecken. Aufgrund der nur geringen Änderung des Bildinhalts aufeinander folgender Bilder lassen sich Merkmalsobjekte sehr viel einfacher zuweisen. Zudem ist der Suchbereich zwischen benachbarten Bildern erheblich kleiner, da die Merkmalsobjekte nur einen geringen Versatz aufweisen. Auf der anderen Seite wurde in den Abschnitten 4.3 und 4.6 bereits auf die Problematik einer geringen Basislinie zwischen zwei Ansichten hingewiesen. Ausgehend von einer statischen Szene, die

[1] Phase Alternating Line
[2] National Television Systems Committee

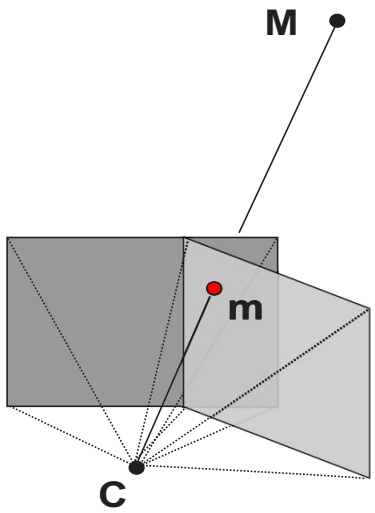


Abbildung 5.1.: Rückprojektion in eine rotierende Kamera: Die Strahlensätze (Triangulation) sind hier nicht anwendbar.

mit einer bewegten Kamera aufgenommen wurde, ist die Basislinie zwischen benachbarten Bildern – wenn überhaupt – sehr gering. Im Falle einer Rotation der Kamera, wie sie häufig in Videosequenzen auftreten kann, ist die Länge der Basislinie sogar null, d.h. die Kamerazentren für zwei benachbarte Bilder sind identisch (siehe Abbildung 5.1). Eine 3D-Rekonstruktion ist in diesem Falle gar nicht möglich, da die Strahlensätze hier keine Anwendung finden. Durch Rauscheinflüsse und durch Quantisierungsfehler führt, wie in Abbildung 4.1 dargestellt, ein zunehmend größer werdender Rekonstruktionsbereich bei Verringerung der Basislinie zu sehr unsicheren Rekonstruktionsergebnissen, d.h. eine 3D-Rekonstruktion im Falle einer translatorischen Kamerabewegung ist für benachbarte Bilder einer Videosequenz kaum möglich.

Eines der Ziele dieses Kapitels ist es, Ansichten innerhalb der Sequenz zu finden, die einen hinreichend großen Abstand der Kamerazentren untereinander aufweisen, um eine 3D-Rekonstruktion zu ermöglichen. Dabei sollen die Korrelationseigenschaften benachbarter Bilder ausgenutzt werden. Ferner sollen die zusätzlichen Informationen, d.h. die Positionen der Merkmalspunkte zwischen diesen Ansichten, ebenfalls für eine Optimierung der Rekonstruktionsergebnisse verwendet werden.

Auf Grundlage bereits gängiger Verfahren zur Rekonstruktion einer dreidimensionalen Szene mithilfe einer bewegten Kamera (engl. *structure-from-motion*, SfM) sind nun folgende Schritte notwendig:

- Verfolgung von Merkmalspunkten innerhalb der Sequenz
- Detektion von Schlüsselbildern
- Initiale Rekonstruktion mit zwei Ansichten
- Hinzufügen von Ansichten zur Verbesserung der Rekonstruktionsergebnisse
- Bestimmung des Kamerapfads
- Optimierung der Ergebnisse mithilfe nichtlinearer Ansätze

5.2. Verfolgung von Merkmalspunkten

5.2.1. Vorbetrachtung

Bei dem in Abschnitt 4.2 vorgestellten Verfahren zur Korrespondenzanalyse wurde bereits die Auswahl geeigneter Merkmalspunkte erläutert. Diese Merkmalspunkte wurden dabei in zwei Bildern getrennt voneinander detektiert und mithilfe eines sehr aufwendigen Verfahrens verglichen, um entsprechende Korrespondenzpunktpaare zu finden. In Videosequenzen ist, wie bereits angedeutet, die Verschiebung eines Punktes von einem Bild zum folgenden Bild der Sequenz sehr gering. Es soll nun gezeigt werden, wie Merkmalspunkte, die ausschließlich im ersten Bild der Sequenz detektiert wurden, im weiteren Verlauf der Sequenz verfolgt werden können (engl. *feature tracking*).

Seien zwei zeitlich benachbarte Bilder einer Sequenz mit I und J gegeben, wobei I das zeitlich vorangegangene Bild darstellt. Ein Punkt $\mathbf{x}$ - konform zur Literatur wird im Folgenden für einen Punkt der Vektor $\mathbf{x}$ anstelle von $\mathbf{m}$ verwendet - innerhalb eines Bildes mit den Koordinaten (x, y) hat dann den Intensitätswert $I(x, y)$ im ersten Bild. Unter Berücksichtigung der zeitlichen Komponente hat dieser Punkt $\mathbf{x}$ an der Position (x, y) zum Zeitpunkt t den Intensitätswert $I(x, y, t)$. Die zeitliche Verschiebung zwischen zwei benachbarten Bildern wird im Weiteren mit τ bezeichnet, so dass die folgende Beziehung gilt:

$$I(x, y, t) = I(x + d_x, y + d_y, t + \tau) \tag{5.1}$$

bzw.

$$I(x, y) = J(x + d_x, y + d_y), \tag{5.2}$$

mit $\mathbf{d} = (d_x, d_y)^T$ als Verschiebungsvektor des Punktes $\mathbf{x} = (x, y)^T$ zwischen den Bildern I und J.

Gl. (5.1) kann aber auch bei einer statischen Szene mit konstanten Lichtverhältnissen nicht als allgemeingültig angesehen werden. Vielmehr werden die Faktoren wie z. B. Verdeckungen oder veränderte Lichtreflexionen aufgrund eines veränderten Betrachtungswinkels der Szene nicht berücksichtigt. Dennoch lässt sich die Bewegung eines Punktes

mit dieser Gleichung sehr gut modellieren, wenn von kleinen Änderungen der Intensitäten und der Vernachlässigung von Verdeckungen ausgegangen wird. Später in diesem Abschnitt wird noch einmal diese Problematik aufgegriffen.

Weiterhin muss noch einmal auf die Mehrdeutigkeit, die schon bei der Korrespondenzanalyse in Abschnitt 4.2 erläutert wurde, zurückgekommen werden. Die Intensitätswerte einzelner Pixel variieren aufgrund von Rauscheinflüssen oder durch Interpolation mit benachbarten Pixeln, bedingt durch das diskrete Sensorfeld der Kamera, in zeitlich benachbarten Bildern. Demnach werden im Weiteren nicht Punkte bzw. einzelne Pixel verfolgt, sondern vielmehr ein Bereich mehrerer Pixel (Fenster) um den zu verfolgenden Punkt. Damit sind aber wieder neue Probleme verknüpft. Zum einen wächst die Wahrscheinlichkeit mit wachsender Fenstergröße, dass Teile des Fensters von Objektkanten verdeckt werden, und zum anderen kann sich der Fensterinhalt infolge von perspektivischen Verzerrungen bzw. durch Rotationen verändern. Eine mögliche Lösung dieser Problematik soll ebenfalls zu einem späteren Zeitpunkt in diesem Abschnitt erläutert werden.

Zunächst soll vorerst wieder eine rein translatorische Verschiebung des Fensters betrachtet werden. Unter Einbeziehung der durch Rauschen und Interpolation auftretenden Störungen lässt sich Gl. (5.1) wie folgt ausdrücken:

$$I(\mathbf{x}) = J(\mathbf{x} + \mathbf{d}) + n(\mathbf{x}), \tag{5.3}$$

wobei der Term n die Rauschstörung angeben soll. Der Verschiebungsvektor $\mathbf{d}$ kann nun über die Minimierung der Fehlerfunktion

$$E = \sum_{\mathbf{x} \in \mathcal{W}} \left[I(\mathbf{x}) - J(\mathbf{x} + \mathbf{d}) \right]^2 \; w \tag{5.4}$$

berechnet werden. $\mathcal{W}$ gibt den Bereich der zu betrachtenden Pixel – und damit das Fenster – an. Der Faktor w entspricht einer Wichtungsfunktion, die im einfachsten Fall z. B. eins sein kann oder aber einer Gaußfunktion entspricht, d.h. Pixel, welche näher an dem zu verfolgenden Punkt (Mittelpunkt des Fensters) liegen, werden stärker gewichtet als Pixel im Randbereich des Fensters. Dies ist vor allem dann sinnvoll, wenn ein großes Fenster verwendet wird. Die Fehlerfunktion ist somit die Summe der quadratischen Abweichungen (engl. *sum of squared differences*, SSD) zwischen den beiden Bildausschnitten.

Wird vorerst von einer kleinen Verschiebung des Fensters zwischen zwei benachbarten Bildern mit geringen Änderungen der Intensitätswerte innerhalb des Fensters ausgegangen, dann sind die Verschiebungen der einzelnen Pixel innerhalb des Fensters näherungsweise konstant. Damit kann die Intensitätsfunktion $J(\mathbf{x} + \mathbf{d})$ durch eine Taylorreihe approximiert und linearisiert werden:

$$J(\mathbf{x} + \mathbf{d}) = J(\mathbf{x}) + \mathbf{g}\mathbf{d}, \tag{5.5}$$

wobei $\mathbf{g} = (J_x, J_y)^T$ die erste Ableitung von J in x- und y-Richtung darstellt. Die Linearisierung erfolgt durch Vernachlässigung der Terme höherer Ordnung. Damit kann die Fehlerfunktion aus Gl. (5.4) angegeben werden mit

$$E = \sum_{\mathbf{x} \in \mathcal{W}} \left[I(\mathbf{x}) - J(\mathbf{x}) - \mathbf{g}\mathbf{d} \right]^2 \, w, \tag{5.6}$$

Die Fehlerfunktion ist somit eine quadratische Funktion in Abhängigkeit von dem Verschiebungsvektor $\mathbf{d}$. Das Minimum der Fehlerfunktion lässt sich über die erste Ableitung bestimmen.

Wird nun Gl. (5.6) nach $\mathbf{d}$ abgeleitet und das Ergebnis zu null gesetzt, so erhält man:

$$\sum_{\mathbf{x} \in \mathcal{W}} \left[I(\mathbf{x}) - J(\mathbf{x}) - \mathbf{g}\mathbf{d} \right] \mathbf{g} \, w = 0. \tag{5.7}$$

Mit $(\mathbf{g}\mathbf{d})\mathbf{g} = (\mathbf{g}\mathbf{g}^T)\mathbf{d}$ und $\mathbf{d}$ als konstanten Verschiebungsvektor innerhalb des Fensters $\mathcal{W}$, kann man Gl. (5.6) wie folgt umstellen:

$$\left(\sum_{\mathbf{x} \in \mathcal{W}} \mathbf{g}\mathbf{g}^T w \right) \mathbf{d} = \sum_{\mathbf{x} \in \mathcal{W}} \left[I(\mathbf{x}) - J(\mathbf{x}) \right] \mathbf{g} \, w. \tag{5.8}$$

Demnach erhält man ein lineares Gleichungssystem mit zwei Gleichungen und zwei Unbekannten, nämlich die Verschiebung $\mathbf{d} = (d_x, d_y)^T$ in x- und y-Richtung. Vereinfacht lässt sich Gl. (5.8) dann ausdrücken durch

$$G\mathbf{d} = \mathbf{e}. \tag{5.9}$$

G ist eine symmetrische 2x2 Matrix mit

$$G = \sum_{\mathbf{x} \in \mathcal{W}} \mathbf{g}\mathbf{g}^T w$$

und $\mathbf{e}$ ein zweidimensionaler Vektor mit

$$\mathbf{e} = \sum_{\mathbf{x} \in \mathcal{W}} \left[I(\mathbf{x}) - J(\mathbf{x}) \right] \mathbf{g} \, w.$$

Für den Fall, dass die Matrix G nicht singulär ist – das ist genau dann der Fall, wenn Bild I um den Punkt $\mathbf{x}$ ausreichend strukturiert ist und damit einen Gradienten in x- und y-Richtung besitzt – ist G invertierbar. Der Verschiebungsvektor $\mathbf{d}$ lässt sich somit durch Berechnung der Gradienten und deren Momente zweiter Ordnung innerhalb eines Bildes und durch die Differenz zwischen den beiden Bildern um den Punkt $\mathbf{x}$ bestimmen.

5.2.2. Affines Bewegungsmodell

Eine Erweiterung, bei der davon ausgegangen wird, dass die Veränderungen von einem Bild zum nächsten nur gering sind, stellt der Feature-Tracker von Shi und Tomasi [165]

dar. Dieser Tracker verwendet das affine Bewegungsmodell, um die Position eines Merkmalspunktes im benachbarten Bild zu finden. Dabei wird nicht nur die Verschiebung eines Punktes berücksichtigt, sondern auch eine Deformation der Umgebung des Punktes infolge von leichten Rotationen und perspektivischen Verzerrungen. Demnach wird nicht nur die Verschiebung des Fensters allein, sondern auch Abweichungen innerhalb des Fensters berücksichtigt, d.h. die Verschiebungen der einzelnen Pixel des Fensters werden nicht mehr als konstant betrachtet. Das affine Bewegungsfeld ist definiert als:

$$\delta = D\mathbf{x} + \mathbf{d} \quad \text{mit} \quad D = \begin{bmatrix} d_{xx} & d_{xy} \\ d_{yx} & d_{yy} \end{bmatrix} \tag{5.10}$$

als Deformationsmatrix und $\mathbf{d}$ als Verschiebungsvektor. Ein Punkt $\mathbf{x}$ in Bild I bewegt sich dann nach $A\mathbf{x} + \mathbf{d}$ im Bild J, mit $A = \underline{I} + D$ und $\underline{I}$ einer 2x2 Einheitsmatrix:

$$J(A\mathbf{x} + \mathbf{d}) = I(\mathbf{x}). \tag{5.11}$$

Da das affine Bewegungsmodell wiederum nur eine Näherung der wahren Deformation des Fensters ist und man davon ausgehen muss, dass die Bilddaten von Rauschen überlagert sind, stellt Gl. (5.11) nur eine Approximation dar. Ziel ist es nun, die Parameter von A und $\mathbf{d}$ zu finden, so dass die Fehlerfunktion

$$E(A, \mathbf{d}) = \sum_{\mathbf{x} \in \mathcal{W}} [I(\mathbf{x}) - J(A\mathbf{x} + \mathbf{d})]^2 w(\mathbf{x}) \tag{5.12}$$

minimal wird.

Um die Fehlerfunktion (5.12) zu minimieren wird sie nach den unbekannten Einträgen der Deformationsmatrix D und des Verschiebungsvektors $\mathbf{d}$ differenziert und das Ergebnis gleich null gesetzt. Durch Linearisierung mittels einer Taylorerweiterung erster Ordnung erhält man

$$J(A\mathbf{x} + \mathbf{d}) = J(\mathbf{x}) + \mathbf{g}^T, \tag{5.13}$$

mit $\mathbf{g} = \left(xI_x, xI_y, yI_x, yI_y, I_x, I_y\right)^T$. Analog den Gleichungen (5.6), (5.7) und (5.8) ergibt sich das folgende lineare System:

$$T\mathbf{z} = \mathbf{a}. \tag{5.14}$$

Der Vektor $\mathbf{z} = \left(d_{xx}, d_{yx}, d_{xy}, d_{yy}, d_x, d_y\right)^T$ enthält die Parameter der Deformationsmatrix D und des Verschiebungsvektors $\mathbf{d}$. Der Fehlervektor

$$\mathbf{a} = \sum_{\mathbf{x} \in \mathcal{W}} [I(\mathbf{x}) - J(\mathbf{x})] \begin{bmatrix} xI_x \\ xI_y \\ yI_x \\ yI_y \\ I_x \\ I_y \end{bmatrix} w(\mathbf{x}) \tag{5.15}$$

hängt von der Differenz der beiden Bilder ab. Die 6×6 Matrix T kann anhand eines Bildes berechnet werden und ist gegeben durch

$$T = \sum_{\mathbf{x} \in \mathcal{W}} \begin{bmatrix} U & V \\ V^T & Z \end{bmatrix} w(\mathbf{x}), \tag{5.16}$$

mit

$$U = \begin{bmatrix} x^2 I_x^2 & x^2 I_x I_y & xy I_x^2 & xy I_x I_y \\ x^2 I_x I_y & x^2 I_y^2 & xy I_x I_y & xy I_y^2 \\ xy I_x^2 & xy I_x I_y & y^2 I_x^2 & y^2 I_x I_y \\ xy I_x I_y & xy I_y^2 & y^2 I_x I_y & y^2 I_y^2 \end{bmatrix}$$

$$V^T = \begin{bmatrix} x I_x^2 & x I_x I_y & y I_x^2 & y I_x I_y \\ x I_x I_y & x I_y^2 & y I_x I_y & y I_y^2 \end{bmatrix}$$

$$Z = \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}.$$

Wegen der Linearisierung wird die Gl. (5.14) nur annähernd erfüllt. Sowohl Gl. (5.14) als auch Gl. (5.9) können mithilfe eines iterativen Verfahrens, wie dem Newton-Raphson-Verfahren, gelöst werden.

Die Wahl eines Bewegungsmodells höherer Ordnung, wie das affine Bewegungsmodell, hat aber nicht nur die Vorteile einer besseren Approximation der eigentlichen Änderung, nämlich die Deformation und die Translation zwischen zwei benachbarter Bilder. Durch die zusätzlich zu bestimmenden Parameter ist die Größe des Fensters entsprechend zu erhöhen. Damit steigt aber auch wieder die Wahrscheinlichkeit, dass ein Fenster, bzw. Teile davon, von Objektkanten verdeckt werden können. Ein weiterer Nachteil entsteht durch periodische Strukturen, die innerhalb größerer Fenster häufiger auftreten können. Bei der Lösung des Minimierungsproblems wächst somit die Gefahr, in ein lokales Minimum zu fallen. Fehler, die dadurch bei der Bestimmung der affinen Deformation entstehen, wirken sich in Gl. (5.14) auch auf die Koordinaten des Verschiebungsvektors $\mathbf{d}$ aus. Eine ähnliche Betrachtung gilt auch für ein perspektivisches Bewegungsmodell, welches Rehg und Witkin [148] vorgeschlagen haben. Aufgrund weiterer zu bestimmender Parameter ist die Fenstergröße gegenüber dem affinen Bewegungsmodell sogar noch einmal zu erhöhen.

Aus den bereits genannten Gründen bezüglich der Fenstergröße soll für die Verfolgung der Merkmalspunkte lediglich das vereinfachte translatorische Bewegungsmodell verwendet werden. Es hat sich gezeigt [17], dass sich dieses für den eigentlichen Verfolgungsprozess hervorragend eignet. Das Fenster $\mathcal{W}$ um den zu verfolgenden Punkt $\mathbf{x}$ hat dabei eine Größe von $(2w_x + 1) \times (2w_y + 1)$. Für w_x und w_y wird typischerweise ein Wert von 1 bis 7 gewählt, so dass sich eine Fenstergröße von 3×3 bis 15×15 ergibt.

Auf das affine Bewegungsmodell wird noch einmal zurückgekommen, wenn es um die Überwachung der verfolgten Merkmalspunkte geht. Bei dieser Überwachung wird die Qualität eines verfolgten Merkmalspunktes nicht zwischen benachbarten Bildern sondern zwischen dem jeweiligen aktuellen Bild und einem zeitlich weiter zurückliegenden Bild ermittelt. Da dann die Bewegung nicht mehr als sehr gering angenommen werden kann, ist ein vereinfachtes translatorisches Bewegungsmodell für den Überwachungsprozess nicht mehr geeignet.

5.2.3. Iteratives Lucas-Kanade-Verfahren

Zur Lösung der Gl. (5.9), bei der nur das translatorische Bewegungsmodell verwendet wurde, soll nun ein iteratives Verfahren angewendet werden. Dieses Verfahren ist in der Literatur unter dem Namen *Lucas-Kanade-Verfahren* [116] oder auch *Lucas-Kanade-Tracker* (LK-Tracker) bekannt. Der LK-Tracker setzt voraus, dass die Bewegung des Punktes von einem Bild zum nächsten nicht zu groß ist. Diese Annahme wurden bereits in der Vorbetrachtung getroffen. Er hat den Vorteil, dass er relativ einfach zu implementieren ist und bei geringem Rechenaufwand sehr gute Ergebnisse liefert. Aus diesem Grund ist der LK-Tracker gerade für die Verfolgung einer großen Anzahl von Merkmalspunkten geeignet.

Ziel des LK-Trackers ist es, den optimalen Verschiebungsvektor $\mathbf{d}_{opt}$ zu bestimmen, der die Gl. (5.9) minimiert. Sei k ein Iterationsindex, der mit 1 für den ersten Durchlauf initialisiert wird. Das Verfahren lässt sich nun wie folgt rekursiv herleiten, indem allgemein für einen Durchgang $k \geq 1$ ein Schätzwert des Verschiebungsvektors aus einem vorherigen Iterationsschritt $\mathbf{d}^{k-1} = \left(d_x^{k-1}, d_y^{k-1}\right)^T$ für den Verschiebungsvektor $\mathbf{d}$ definiert werden kann. Es sei J_k das neue Bild, welches um den Wert $\mathbf{d}^{k-1}$ verschoben wurde:

$$J_k(\mathbf{x}) = J(\mathbf{x} + \mathbf{d}^{k-1}). \tag{5.17}$$

Definiert man nun den verbleibenden Rest der Verschiebung $\eta^k = \left(\eta_x^k, \eta_y^k\right)^T$, welcher sich aus der Differenz zwischen dem optimalen Verschiebungsvektor $\mathbf{d}_{opt}$ und dem Verschiebungsvektor $\mathbf{d}^{k-1}$ aus dem Iterationsschritt $k-1$ ergibt. Mit Gl. (5.17) und η^k ergibt sich die Fehlerfunktion

$$E(\eta^k) = \sum_{\mathbf{x} \in \mathcal{W}} \left[I(x,y) - J_k(x + \eta_x^k, y + \eta_y^k) \right]^2, \tag{5.18}$$

welche es zu minimieren gilt. Die Wichtungsfunktion w aus Gl. (5.4) wurde zu 1 gesetzt, da für das translatorische Bewegungsmodell kleine Fenster zur Verfolgung verwendet werden.

Der Verschiebungsrest η^k kann entsprechend Gl. (5.9) ermittelt werden mit

$$\eta^k = G^{-1}\mathbf{e}^k. \tag{5.19}$$

Hierbei ist $\mathbf{e}^k$ – welcher auch als Fehlversatzvektor bezeichnet wird – wie folgt definiert:

$$\mathbf{e}^k = \sum_{\mathbf{x} \in \mathcal{W}} \begin{bmatrix} \Delta I_k(\mathbf{x}) I_x(\mathbf{x}) \\ \Delta I_k(\mathbf{x}) I_y(\mathbf{x}) \end{bmatrix} \tag{5.20}$$

mit

$$\Delta I_k(\mathbf{x}) = I(\mathbf{x}) - J_k(\mathbf{x})$$

und

$$\nabla I = \begin{pmatrix} I_x(\mathbf{x}) \\ I_y(\mathbf{x}) \end{pmatrix} = \left(\frac{\partial I(\mathbf{x})}{\partial x}, \frac{\partial I(\mathbf{x})}{\partial y} \right)^T.$$

Der Vorteil in dieser rekursiven Berechnung ist, dass die lokalen Ableitungen I_x und I_y nur einmal zu Beginn berechnet werden müssen und während der Iterationen konstant bleiben. Dadurch ist auch die 2×2 Matrix G konstant für die gesamten Rechenoperationen. Nur der Fehlversatzvektor $\mathbf{e}^k$, welcher den verbleibenden Versatz zwischen den Bildausschnitten nach der Verschiebung um $\mathbf{d}^{k-1}$ angibt, muss in jeder Iteration k neu ermittelt werden. Nachdem η^k bestimmt wurde, ergibt sich für den k-ten Iterationsschritt der neue Verschiebungsvektor zu

$$\mathbf{d}^k = \mathbf{d}^{k-1} + \eta^k. \tag{5.21}$$

Die gesamte Berechnung wird solange wiederholt, bis der verbleibende Versatz η^k unter einen Schwellwert τ (in [17] wird ein Wert von 0.03 Pixel vorgeschlagen) fällt oder eine maximale Anzahl von Durchläufen (z. B. 20 Iterationen nach [17]) erreicht ist. Es hat sich gezeigt, dass im Durchschnitt 5 Iterationen ausreichend sind, um die Konvergenz zu erreichen.

Für die erste Iteration ($k = 1$) wird der Verschiebungsvektor mit

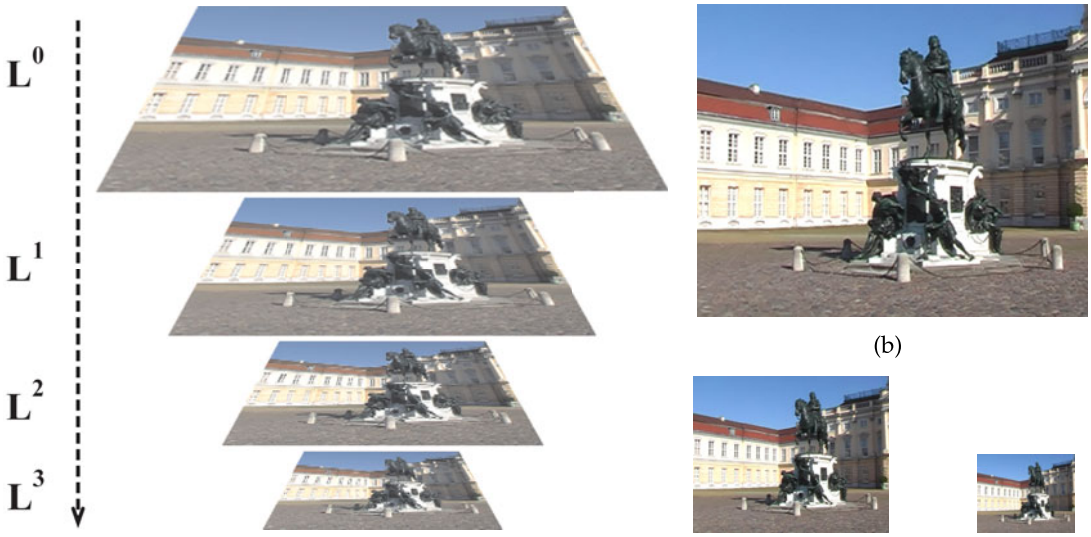
$$\mathbf{d}^0 = \mathbf{0}$$

initialisiert. Sei K die Anzahl der Iterationen, die benötigt werden, damit $\mathbf{d}^k$ gegen die Lösung konvergiert, dann ergibt sich für den Verschiebungsvektor $\mathbf{d}$

$$\mathbf{d} = \mathbf{d}^K = \sum_{k=1}^{K} \eta^k. \tag{5.22}$$

Dieser Vektor minimiert Gl. (5.4).

Mit dem iterativen Lucas-Kanade-Verfahren wurde unter Ausnutzung des translatorischen Bewegungsmodells mit kleinen Fenstern um die zu verfolgenden Merkmalspunkte ein Verfahren dargelegt, welches kleine Bewegungen bzw. Veränderungen des Bildinhalts erfassen kann. Bei stärkeren Bewegungen, z. B. verursacht durch eine schnelle Kameraführung oder durch nahe gelegene Vordergrundobjekte, müssten größere Fenster gewählt



(a) Interpretation der pyramidenförmigen Repräsentation (b) (c) (d)

Abbildung 5.2.: Pyramidenförmige Repräsentation: (a) Interpretation, (b) ursprüngliches Bild mit voller Auflösung, (c) bzw. (d) reduzierte Auflösung um den Faktor 2 bzw. 4

werden. Die Nachteile größerer Fenster wurden bereits genannt. Eine Erweiterung des Lucas-Kanade-Verfahrens, welche auch stärkere Bewegungen erfassen kann, soll nun unter Beibehaltung kleiner Fenster in Form einer pyramidenförmigen Struktur erläutert werden.

5.2.4. Pyramidenförmige Implementierung des iterativen Lucas-Kanade-Verfahrens

Eine Erweiterung des LK-Trackers stellt die pyramidenförmige Implementierung des iterativen Lucas-Kanade-Verfahrens dar [17], indem zuerst verschiedene Auflösungsstufen der Bilder erzeugt werden. Das Lucas-Kanade-Verfahren wird hierbei auf alle Auflösungsstufen angewendet, beginnend mit der Stufe der niedrigsten Bildauflösung.

Sei $I = I^0$ das Bild mit der ursprünglichen Auflösung von $n_x \times n_y$, wobei $n_x = n_x^0$ die Breite und $n_y = n_y^0$ die Höhe des Bildes ist. Die weiteren Auflösungsstufen werden rekursiv aus vorangegangenen Stufen berechnet. Um Aliasing-Effekte bei der Reduzierung der Auflösung zu vermeiden, soll dabei gleichzeitig eine Tiefpass-Filterung angewendet werden. Seien nun $L = \{0, 1, ..., L_m\}$ die Stufen der pyramidenförmigen Repräsentation,

dann ist, allgemein betrachtet, I^L das Bild der Stufe L mit der Auflösung $n_x^L \times n_x^L$ und lässt sich durch Interpolation benachbarter Pixel (Tiefpass) berechnen mit

$$\begin{aligned} I^L(x,y) = & \frac{1}{4} I^{L-1}(2x,2y) + \cdots \\ & \frac{1}{8}\Big(I^{L-1}(2x-1,2y) + I^{L-1}(2x+1,2y) + \cdots \\ & I^{L-1}(2x,2y-1) + I^{L-1}(2x,2y+1)\Big) + \cdots \\ & \frac{1}{16}\Big(I^{L-1}(2x-1,2y-1) + I^{L-1}(2x+1,2y-1) + \cdots \\ & I^{L-1}(2x-1,2y+1) + I^{L-1}(2x+1,2y+1)\Big). \end{aligned} \tag{5.23}$$

Die Auflösung der Stufe L hat sich somit gegenüber der Stufe $L-1$ um die Hälfte reduziert.

Unter Verwendung der Gl. (5.23) kann nun die pyramidenförmige Repräsentation der Bilder I und J mit $\left\{I^L\right\}_{L=0,\ldots,L_m}$ und $\left\{J^L\right\}_{L=0,\ldots,L_m}$ bestimmt werden. L_m entspricht der höchsten Stufe in der Pyramide. Typische Werte für L_m sind, je nach Auflösung der Bilder in der Videosequenz, Werte von 1 bis 4. In Abbildung (5.2) ist die pyramidenförmige Repräsentation graphisch dargestellt.

Beispiel: Bei einer Standardauflösung von 720×480 Bildpunkten im NTSC-Format ergibt sich für die einzelnen Pyramidenstufen I^1, I^2, I^3 und I^4 eine Auflösung von 360×240, 180×120, 90×60 und 45×30 Bildpunkten. Die Wahl von mehr als vier Stufen wäre hier wenig ratsam, da in den unteren Stufen kaum noch Details enthalten sind. Auch für Auflösungen im HDTV-Format, wie z. B. 1920x1080 (hier ist die Auflösung für I^4 120x68 Bildpunkte), ist eine Stufenanzahl weit größer als vier nicht zu empfehlen, da nur noch grobe Details erkennbar bleiben.

Bezeichnen man den zu verfolgenden Merkmalspunkt im Ausgangsbild I mit $\mathbf{x}$, dann lässt sich seine Position in den entsprechenden Auflösungsstufen berechnen mit

$$\mathbf{x}^L = \frac{\mathbf{x}}{2^L} \quad \text{mit} \quad \mathbf{x}^0 = \mathbf{x}. \tag{5.24}$$

Ausgehend von der höchsten Stufe der Pyramide L_m wird der initiale Verschiebungsvektor mit $\nu^{L_m} = \left(\nu_x^{L_m}, \nu_y^{L_m}\right)^T = (0,0)^T$ definiert. Verfährt man allgemeiner und bezeichnet ν^L als initialen Verschiebungsvektor für die Stufe L, der in Stufe $L+1$ ermittelt wurde, dann lässt sich der tatsächliche Verschiebungsvektor in Stufe L über den Verschiebungsrest $\mathbf{d}^L = \left(d_x^L, d_y^L\right)^T$ durch Minimierung der Fehlerfunktion E^L ermitteln:

$$E(\mathbf{d}^L) = \sum_{\mathbf{x} \in \mathcal{W}} \left(I(x,y) - J(x + \nu_x^L + d_x^L, y + \nu_y^L + d_y^L)\right)^2. \tag{5.25}$$

Die Fenstergröße wird in allen Stufen konstant gehalten. Da ν^L bereits als initiale Schätzung für die Verschiebung angenommen wird, ist der Verschiebungsrest $\mathbf{d}^L$ klein und

kann durch das iterative Lucas-Kanade-Verfahren berechnet werden. Nachdem der Verschiebungsrest $\mathbf{d}^L$ ermittelt wurde, wird das Ergebnis der Berechnung an die nächste Stufe $L-1$ weitergegeben. Als neuer Schätzvektor ν^{L-1} ergibt sich

$$\nu^{L-1} = 2(\nu^L + \mathbf{d}^L). \tag{5.26}$$

Der nächste Verschiebungsrest $\mathbf{d}^{L-1}$ wird daraufhin wieder nach dem Lucas-Kanade-Verfahren berechnet, bis das Bild mit der höchsten Auflösung ($L = 0$) erreicht ist. Der resultierende Verschiebungsvektor $\mathbf{d}$, welcher die Fehlerfunktion in Gl. (5.4) minimiert, ist:

$$\mathbf{d} = \mathbf{d}^0 + \nu^0. \tag{5.27}$$

Dieser Verschiebungsvektor kann auch direkt durch die Verschiebungsreste aus jeder Stufe angeben werden und lässt sich wie folgt herleiten. Ausgehend von der höchsten Stufe L_m ist $\nu^{L_m} = (0,0)^T$. Mit Gl. (5.26) ergibt sich für die folgenden Stufen:

$$\begin{aligned}
\nu^{L_m-1} &= 2 \cdot d^{L_m} \\
\nu^{L_m-2} &= 2\left(2 \cdot d^{L_m} + d^{L_m-1}\right) \\
\nu^{L_m-3} &= 2\left[2\left(2 \cdot d^{L_m} + d^{L_m-1}\right) + d^{L_m-2}\right] \\
&\vdots = \vdots \\
\nu^{L} &= \sum_{l=L+1}^{L_m} 2^l \mathbf{d}^l \\
&\vdots = \vdots \\
\nu^{0} &= \sum_{l=1}^{L_m} 2^l \mathbf{d}^l
\end{aligned} \tag{5.28}$$

Setzt man das Ergebnis von Gl. (5.28) in Gl. (5.27) ein, so erhält man

$$\mathbf{d} = \sum_{l=0}^{L_m} 2^l \mathbf{d}^l. \tag{5.29}$$

Mit Gl. (5.29) lässt sich auch die maximale Verschiebung bzw. der Verschiebungsgewinn durch Verwendung der pyramidenförmigen Struktur ermitteln. Sei $d_{max,LK}$ die maximal zulässige Verschiebung für das allgemeine Lucas-Kanade-Verfahren, dann ist $d_{max} = (2^{L_m+1} - 1)d_{max,LK}$ die maximal zulässige Verschiebung bei Verwendung der Auflösungspyramide. Damit ergibt sich ein Verschiebungsgewinn $\mathcal{G}_V = (2^{L_m+1} - 1)$.

Beispiel: Bei einer Anzahl von fünf Stufen ($L_m = 4$) für die Auflösungspyramide und einer maximal zulässigen Pixelverschiebung von $d_{max,LK} = 2$ für das Lucas-Kanade-Verfahren wäre die maximal zulässige Verschiebung $d_{max} = (2^{4+1} - 1) \cdot 2 = 62$, bzw. der Verschiebungsgewinn $\mathcal{G}_V = 31$.

Ziel: Lokalisierung des Punktes $\mathbf{x}$ aus Bild I im Bild J.

- *Erzeuge die Auflösungspyramide für I und J*: $\{I^L\}_{L=0,\ldots,L_m}$ und $\{J^L\}_{L=0,\ldots,L_m}$
- *Initialisierung des Schätzwertes*: $\nu^{L_m} = \left(\nu_x^{L_m}, \nu_y^{L_m}\right)^T = (0,0)^T$
  Wiederhole für $L = L_m$ bis 0, mit einer Schrittweite von -1
  1. *Position des Punktes $\mathbf{x}$ im Bild I*: $\mathbf{x}^L = \frac{\mathbf{x}}{2^L}$
  2. *Ableitung von I^L in x-Richtung*: $I_x(x,y) = \frac{I^L(x+1,y) - I^L(x-1,y)}{2}$
  3. *Ableitung von I^L in y-Richtung*: $I_y(x,y) = \frac{I^L(x,y+1) - I^L(x,y-1)}{2}$
  4. *Berechnung der lokalen Gradientenmatrix*: $G = \sum_{\mathbf{x}\in\mathcal{W}} \begin{bmatrix} I_x^2(\mathbf{x}) & I_x(\mathbf{x})I_y(\mathbf{x}) \\ I_x(\mathbf{x})I_y(\mathbf{x}) & I_y^2(\mathbf{x}) \end{bmatrix}$
  5. *Initialisierung des iterativen Lucas-Kanade-Verfahrens*: $\mathbf{d}^0 = (0,0)^T$
     Für $k = 1$ bis K, mit einer Schrittweite von 1 (oder bis $\|\eta^k\| < \tau$)
     a) *Bilddifferenz*: $\Delta I_k(\mathbf{x}) = I^L(\mathbf{x}) - J^L(\mathbf{x} + \nu^L + \mathbf{d}^{k-1})$
     b) *Fehlversatzvektor*: $\mathbf{e}^k = \sum_{\mathbf{x}\in\mathcal{W}} \begin{bmatrix} \Delta I_k(\mathbf{x}) I_x(\mathbf{x}) \\ \Delta I_k(\mathbf{x}) I_y(\mathbf{x}) \end{bmatrix}$
     c) *Verschiebungsrest*: $\eta^k = G^{-1}\mathbf{e}^k$
     d) *Schätzwert für den nächsten Durchgang*: $\mathbf{d}^k = \mathbf{d}^{k-1} + \eta^k$
     Ende der Schleife über k
  6. *Verschiebungsvektor der Stufe L*: $\mathbf{d}^L = \mathbf{d}^k$
  7. *Schätzwert für die nächste Stufe $L-1$*: $\nu^{L-1} = 2(\nu^L + \mathbf{d}^L)$

  Ende der Schleife über L
- *Resultierender Verschiebungsvektor (Stufe $L = 0$)*: $\mathbf{d} = \nu^0 + \mathbf{d}^0$
- *Position des Punktes im Bild J*: $\mathbf{x}' = \mathbf{x} + \mathbf{d}$

Ergebnis: Die korrespondierende Position des Punktes $\mathbf{x}$ in Bild J.

Algorithmus 5: Zusammenfassung der Berechnungsschritte des pyramidenförmigen LK-Trackers

In Algorithmus 5 sind die verschiedenen Rechenschritte zur Bestimmung des Verschiebungsvektors noch einmal zusammengefasst.

Bei den bisherigen Überlegungen wurde lediglich die Verschiebung eines Punktes von einem Bild der Videosequenz zum folgenden Bild betrachtet. In Algorithmus 5 wurde ein Startwert des Verschiebungsvektors $\nu^{L_m} = \left(\nu_x^{L_m}, \nu_y^{L_m}\right)^T = (0,0)^T$ initialisiert. Die Verschiebung des Punktes kann aber auch aus der Bewegung des Punktes in den vorangegangenen Bildern der Sequenz geschätzt werden und somit als Initialisierung für das nächste Bild dienen. Im Allgemeinen kann davon ausgegangen werden, dass sich in einer kontinuierlichen Szene die Bewegungsrichtung und -weite eines Punktes nicht schlagartig ändert. Als Initialisierung könnte im einfachsten Fall der Verschiebungsvektor aus den Berechnungen in den vorherigen beiden Bildern der Sequenz dienen. Eine genauere Schätzung der

Bewegung erhält man z. B. durch die Verwendung eines Kalman-Filters [91], da dieser – durch Berücksichtigung mehrerer vorangegangener Bilder – auch die Veränderung in der Bewegung mit erfassen kann.

Berechnung der Verschiebung mit Subpixelgenauigkeit.

Die Konvergenz von Gleichung 5.22 kann nur dann erreicht werden, wenn die Berechnung nicht nur mittels ganzzahliger Pixelpositionen durchgeführt wird, sondern auch Positionen zwischen den Pixeln betrachtet werden. Es soll an dieser Stelle eine Lösung gezeigt werden, wie mittels bipolarer Interpolation solche Zwischenwerte berechnet werden können.

Sei L die Pyramidenstufe, für die der Wert an der Stelle $I^L(x, y)$ benötigt wird, wobei x und y keine ganzen Zahlen sind. Seien nun x_0 und y_0 die ganzzahligen Anteile von x und y, dann bilden α_x und α_y den Rest, so dass gilt

$$x = x_0 + \alpha_x \tag{5.30}$$

$$y = y_0 + \alpha_y. \tag{5.31}$$

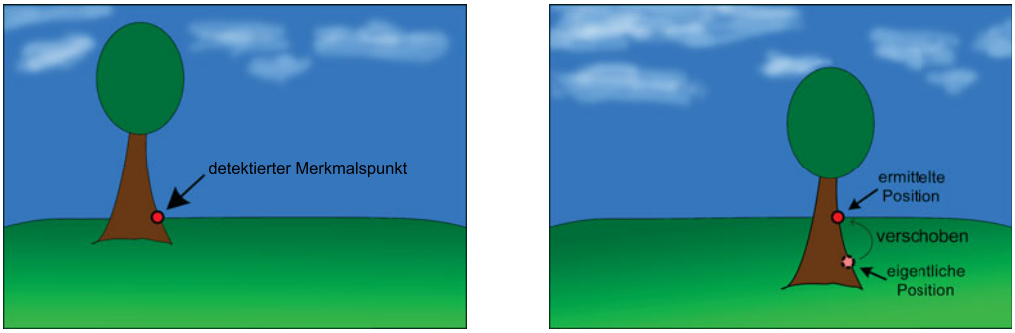
Der Intensitätswert an der Stelle $I^L(x, y)$ kann somit mittels bipolarer Interpolation aus den benachbarten Werten berechnet werden mit

$$\begin{aligned} I^L(x,y) =& (1-\alpha_x)(1-\alpha_y)I^L(x_0,y_0) + \alpha_x(1-\alpha_y)I^L(x_0+1,y_0) + \\ & (1-\alpha_x)\alpha_y I^L(x_0,y_0-1) + \alpha_x\alpha_y I^L(x_0+1,y_0+1). \end{aligned} \tag{5.32}$$

Verfolgung von Merkmalspunkten im Randbereich.

Bei der Verfolgung von Merkmalspunkten, z. B. mithilfe eines LK-Trackers, sollen diese auch im Randbereich eines Bildes weiter verfolgt werden können. Der Randbereich stellt dabei ein spezielles Problem dar, da ein Teil des Fensters um einen Merkmalspunkt außerhalb des Bildes liegen kann. Vor allem bei Verwendung einer Auflösungspyramide mit einer großen Stufenanzahl $L_m + 1$ treten solche Fälle verstärkt auf. Wenn gefordert wird, dass sich immer der vollständige Bereich des Fensters $\mathcal{W}$ der Größe $(2w_x + 1) \times (2w_y + 1)$ im Bild befinden muss, dann existiert ein Randbereich des Bildes der Größe w_x bzw. w_y, in dem Merkmalspunkte nicht weiter verfolgt werden können. Sei L_m die Höhe der Pyramide, dann erhält man einen effektiven Randbereich im Originalbild I der Größe $2^{L_m} w_x$ bzw. $2^{L_m} w_y$.

Beispiel: Seien die Fenstergröße mit $w_x = w_y = 5$ und die Höhe der Auflösungspyramide $L_m = 3$ gegeben, so ergibt sich ein effektiver Randbereich des Bildes, in dem Merkmalspunkte nicht weiter verfolgt werden, von 40 Pixeln in der Auflösungsstufe $L = 0$.



(a) (b)

Abbildung 5.3.: Die richtige Position eines Punktes an Tiefensprüngen kann nur schwer ermittelt werden, wenn der Hintergrund z. B. sehr homogene Flächen aufweist. Hier wurde ein Punkt an der Schnittstelle des Baumstammes mit dem Übergang zum Horizont detektiert. Durch die Veränderung der Position des Baumes im Bild (b) gegenüber dem Bild (a), repräsentiert der Punkt nicht mehr die selbe Position im Raum (Quelle: [193]).

Aufgrund des erhöhten Randbereichs bei Verwendung einer Auflösungspyramide ist es ratsam, das Verfahren zur Verfolgung von Merkmalspunkten dahingehend zu erweitern, dass Punkte, deren Fenster teilweise außerhalb des Bildes liegen, weiterhin verfolgt werden können. Für die notwendigen Rechenschritte sind ausschließlich die Werte des Fensters zu berücksichtigen, die innerhalb des Bildbereichs liegen. Dabei gilt es zu beachten, dass sich der Summationsbereich während der Iterationsschritte des Lucas-Kanade-Verfahrens verändern kann. Zum Beispiel ergibt sich ein neuer Wert für die Bilddifferenz ΔI_k, wenn sich der Verschiebungsvektor $\left(v_x^L + d_x^{k-1}, v_y^L + d_y^{k-1}\right)^T$ verändert. Bei der Berechnung des Fehlversatzvektors $\mathbf{e}^k$ und der Gradientenmatrix G ist zu beachten, dass die Summationsbereiche identisch sein müssen, d.h. G muss ggf. in jedem Iterationsschritt neu bestimmt werden.

Liegt allerdings nicht nur ein Teil des Fensterbereichs außerhalb des Bildes sondern auch der eigentlich zu verfolgende Merkmalspunkt, so wird dieser verworfen und nicht weiter verfolgt. Weiterführende Implementierungsdetails sind in [193] zusammengefasst.

5.2.5. Qualität der verfolgten Merkmalspunkte

Ein wichtiger Prozess bei der Verfolgung von Merkmalspunkten stellt die Überprüfung der Qualität bzw. die Überwachung der verfolgten Merkmalspunkte dar. Bislang wurde bei der Verfolgung von einer rein translatorischen Bewegungen zwischen benachbarten Bildern einer Videosequenz ausgegangen. Zudem wurde bereits ein Fall erwähnt, bei dem Merkmalspunkte verworfen werden, wenn sie den Bildbereich verlassen. Es gibt aber weitere Faktoren, die dazu führen können, dass die eigentliche Position eines Merkmalspunktes nicht mehr korrekt bestimmt werden kann.

Zum Beispiel kann es vorkommen, dass ein Merkmalspunkt im Verlauf der Videosequenz von einem anderen Objekt in der Szene verdeckt wird oder dass bei der Auswahl eines Merkmalspunktes dieser direkt auf einer Objektkante liegt, wie in Abbildung 5.3 dargestellt ist. Eine einfache Verfolgung solch eines Punktes mit einem LK-Tracker, wie er im vorangegangenen Abschnitt beschrieben wurde, wäre immer noch möglich, da sich das Minimum der Fehlerfunktion aus Gl. (5.4) finden ließe.

Um das Ergebnis der Lokalisierung eines Mermalspunktes ggf. zu verwerfen und damit die Verfolgung des Punktes zu beenden, wird, analog zu Gl. (5.4), eine Fehlerfunktion definiert, welche die Summe der quadratischen Abweichungen (SSD) zwischen dem aktuellen Bild J der Sequenz und dem Bild I^n, in dem der Punkt detektiert wurde, berechnet. Mit Zunahme des Abstands der beiden Bilder J und I^n lässt sich die Verschiebung der Bildbereiche nur schlecht über ein rein translatorisches Bewegungsmodell bestimmen, weshalb an dieser Stelle nun das affine Bewegungsmodell verwendet werden soll. Sei $\mathbf{d}$ der Verschiebungsvektor des Merkmalspunktes $\mathbf{x} = (x, y)$ zwischen Bild I^n und Bild J, dann ergibt sich für die Fehlerfunktion entsprechend Gl. (5.12):

$$E(A, \mathbf{d}) = SSD = \sum_{\mathbf{x} \in \mathcal{W}_Q} \left(J(A\mathbf{x} + \mathbf{d}) - I^n(\mathbf{x})\right)^2 . \tag{5.33}$$

Die affinen Transformationsparameter in Matrix A können durch Lösen der Gleichung (5.14) ermittelt werden. Aufgrund der zusätzlichen Parameter sollte die Fenstergröße $\mathcal{W}_Q$ in Gl. (5.33) größer gewählt werden als bei dem LK-Tracker.

Die Entscheidung, ob ein Merkmalspunkt weiter verfolgt werden soll, wird mittels eines Schwellwertes τ getroffen. Solange der Fehler kleiner als der Schwellwert ist ($SSD \leq \tau$), ist die Qualität des verfolgten Punktes ausreichend. Die Herausforderung hierbei ist, auf welchen Wert der Schwellwert festgelegt wird. Tomasi und Kanade verwenden in [180] einen nicht bei Null startenden, linear ansteigenden Schwellwert, da der SSD-Wert für alle Punkte von Bild zu Bild zunimmt.

Bestimmung des Schwellwertes

In [53] wird von Fusiello et al. eine Methode - auch bekannt unter dem Namen *X84-Rejection-Rule* - zur robusten und automatischen Bestimmung des Schwellwertes aus allen berechneten SSD-Werten vorgestellt. Bei ihnen wird der Schwellwert anhand des *Medianwertes der absoluten Abweichungen* (engl. *median absolute deviation*, **MAD**) ermittelt. Der MAD-Wert ist gegeben durch

$$MAD = \underset{i}{\text{mad}} \, \|e_i - \underset{j}{\text{median}} \, e_j\|. \tag{5.34}$$

Dabei ist e_i bzw. e_j der SSD-Wert für den Merkmalspunkt i bzw. j zwischen dem aktuellen Bild und dem ersten Bild, in dem der Punkt detektiert wurde. Ein Merkmalspunkt gilt

als „verloren“, d.h. wird nicht weiter verfolgt, wenn er einen SSD-Wert von mehr als den k-fachen MAD-Wert aufweist. Ausgehend von normalverteilten SSD-Werten wurde in [53] für $k = 5,2$ ermittelt.

Sie erweitern auch die Fehlerfunktion aus Gl. (5.33) zu einer „normierten“ SSD, um Helligkeitunterschiede zwischen den Bildern auszugleichen. Ihre Fehlerfunktion ist definiert als

$$E(A, \mathbf{d}) = SSD_{norm} = \sum_{\mathbf{x} \in W_Q} \left(\frac{J(A\mathbf{x} + \mathbf{d}) - \mu_J}{\sigma_J} - \frac{I^n(\mathbf{x}) - \mu_I}{\sigma_I} \right)^2 . \tag{5.35}$$

Hierbei sind μ_J und μ_I die Mittelwerte und σ_J und σ_I die Standardabweichungen der Intensitätswerte der Bilder J und I^n.

In [61] beschreiben Hager und Belhumuer sowie Ma et al. in [119] einen SSD-basierten Tracker, welcher schon bei der Bestimmung der Transformation die Veränderung in der Helligkeit kompensiert. In [89] werden bei der Verfolgung der Merkmalspunkte auch Veränderungen in der Beleuchtung sowie Reflektionen berücksichtigt.

Normierte Kreuzkorrelation

Anstelle einer auf der SSD aufbauenden Fehlerfunktion kann eine Fehlerfunktion verwendet werden, die auf der *normierten Kreuzkorrelation* (engl. *normalized cross correlation*, NCC), welche Werte zwischen -1 und 1 annehmen kann, basiert. Der Vorteil einer NCC-basierten Fehlerfunktion wurde bereits in Abschnitt 4.2 erläutert. Sie ist für einen Merkmalspunkt $\mathbf{x} = (x, y)^T$ im Bild I^n, dessen Verschiebung $\mathbf{d} = \left(d_x, d_y\right)^T$ zum aktuellen Bild J und der affinen Transformationsmatrix A, wie folgt definiert:

$$E(A, \mathbf{d}) = \mathrm{NCC}\,(A, \mathbf{d}) = \frac{\sum\limits_{\mathbf{x} \in \mathcal{W}_Q} (I(\mathbf{x}) - \mu_I)(J(A\mathbf{x} + \mathbf{d}) - \mu_J)}{\sqrt{\sum\limits_{\mathbf{x} \in \mathcal{W}_Q} (I(\mathbf{x}) - \mu_I)^2 \sum\limits_{\mathbf{x} \in \mathcal{W}_Q} (J(A\mathbf{x} + \mathbf{d}) - \mu_J)^2}}. \tag{5.36}$$

Für $\mathcal{W}_Q$ gelten die selben Bedingungen wie bei Gl. 5.33. Als Größe für den Schwellwert τ werden im Allgemeinen Werte zwischen 0,7 und 0,9 gewählt. Im Laufe der Untersuchungen hat sich gezeigt, dass ein Schwellwert von 0,8 gute Ergebnisse liefert, d.h. Merkmalspunkte, deren Wert $\mathrm{NCC}\,(A, \mathbf{d}) \geq \tau = 0,8$ werden weiter verfolgt.

Reduzierung der Rechenkomplexität

Die Berechnung der affinen Transformationsparameter ist bei einer großen Anzahl zu verfolgender Merkmalspunkte ein äußerst rechenintensiver Prozess. Während bei dem LK-Tracker mit kleiner Fenstergröße lediglich die beiden Translationsparameter berechnet wurden, müssen bei der Qualitätsüberprüfung vier Parameter der Deformationsmatrix mithilfe größerer Fenster ermittelt werden. D.h. der Großteil der Rechenoperation

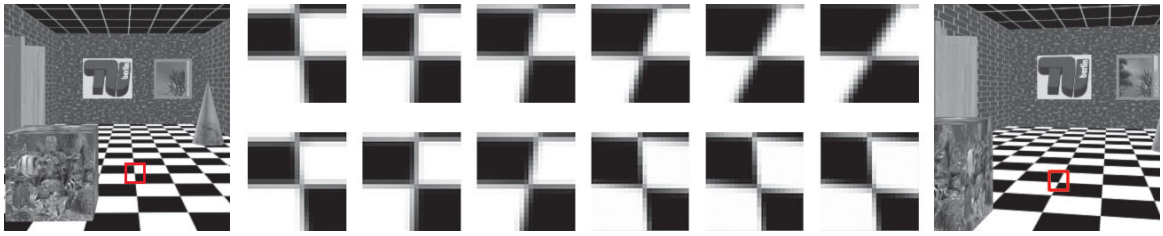


Abbildung 5.4.: Qualitätsbestimmung eines Merkmalspunktes. Oben: rein translatorisches Bewegungsmodell, unten: kombiniertes translatorisches und affines Bewegungsmodell.

liegt nicht in der eigentlichen Verfolgung der Punkte sondern in der Überwachung derselbigen. Abbildung 5.4 zeigt ein Beispiel für die beiden verwendeten Bewegungsmodelle bei der Verfolgung (rein translatorisches Bewegungsmodell) und der Qualtitätsprüfung (kombiniertes translatorisches und affines Bewegungsmodell).

Infolge der Untersuchungen bezüglich der Qualitätsüberprüfung haben sich zwei interessante Eigenschaften ergeben, durch die die Rechenkomplexität des Überwachungsprozesses erheblich reduziert werden kann. Zum einen hat sich herausgestellt, dass die normierte Kreuzkorrelation bei Verwendung des rein translatorischen Bewegungsmodells als Gütekriterium sehr aussagekräftig ist, d.h. die affinen Transformationsparameter in Gl. (5.36) werden vorerst nicht berechnet. Somit wird für die Matrix A die Einheitsmatrix $\underline{I}$ eingesetzt ($A = \underline{I}$). Merkmalspunkte, deren NCC-Wert größer als der verwendete Schwellwert τ ist, würden auch mit Berücksichtigung der affinen Transformationsparameter oberhalb dieser Schwelle liegen. Zum anderen hat sich ergeben, dass Merkmalspunkte, deren NCC-Wert sehr viel kleiner als der verwendete Schwellwert τ ist, auch durch die Berechnung der Deformationsmatrix nicht über diesen Schwellwert gelangen und demnach nicht weiter verfolgt werden. Auch in diesem Fall wäre die Berechnung der affinen Transformationsparameter überflüssig und würde die Rechenzeit unnütz erhöhen. Für den Fall, dass der NCC-Wert eines Merkmalspunktes allerdings nur knapp unterhalb der Schwelle liegt, erweist sich die Berechnung der Deformationsmatrix als durchaus sinnvoll, da hier die Änderungen des Bildausschnitts um den Merkmalspunkt durch die affinen Transformationsparameter besser modelliert werden können und somit den NCC-Wert über die Schwelle τ anheben kann.

Anhand dieser Betrachtungen scheint es also sinnvoll, zwei Schwellwerte zu definieren, zwischen denen die Berechnung der affinen Transformationsparameter vorgenommen wird. Bezeichnet man den oberen Schwellwert als τ_{gut}, d.h. Merkmalspunkte, deren NCC($\mathbf{d}$)-Wert oberhalb von τ_{gut} liegt, werden als „gut“ deklariert und können weiter verfolgt werden, und den unteren Schwellwert als $\tau_{schlecht}$, d.h. Merkmalspunkte, deren NCC($\mathbf{d}$)-Wert unterhalb von $\tau_{schlecht}$ liegt, werden als „schlecht“ deklariert und werden

Ziel: Identifikation von Merkmalspunkten, die nicht weiterverfolgt werden sollen, da sie verdeckt wurden oder sich zu stark verändert haben.

Wiederhole für alle Merkmalspunkte i

1. Bestimme für die Merkmalspunkte i den Fehler e_i zwischen dem aktuellen Bild J und dem Bild I^n mittels Gl. (5.33) oder Gl. (5.36) mit $A = \underline{I}$.
   a) Wenn $e_i \leq \tau_{gut}$ (für SSD) bzw. $e_i \geq \tau_{gut}$ (für NCC), wird der Punkt als „gut“ deklariert und weiter verfolgt.
   b) Wenn $e_i > \tau_{schlecht}$ (für SSD) bzw. $e_i < \tau_{schlecht}$ (für NCC), wird der Punkt als „verloren“ deklariert und nicht weiter verfolgt.
2. Ansonsten bestimme die affinen Transformationsparameter mit Gl. (5.14).
3. Bestimme erneut den Fehler e_i unter Verwendung der affinen Transformationparameter ($A = D + \underline{I}$).
   a) Wenn $e_i \leq \tau_{gut}$ (für SSD) bzw. $e_i \geq \tau_{gut}$ (für NCC), wird der Punkt als „gut“ deklariert und weiter verfolgt.
   b) Wenn $e_i > \tau_{gut}$ (für SSD) bzw. $e_i < \tau_{gut}$ (für NCC), wird der Punkt als „verloren“ deklariert und nicht weiter verfolgt.

Ende der Schleife

Algorithmus 6: Bestimmung der Qualität von Merkmalspunkten

nicht weiter verfolgt. Für NCC($\mathbf{d}$)-Werte zwischen diesen beiden Schwellen wird die Deformationsmatrix berechnet und der NCC(A, $\mathbf{d}$)-Wert nach Gl. (5.36) bestimmt. Liegt dieser dann oberhalb von τ_{gut}, wird der Punkt weiter verfolgt; bleibt er allerdings unterhalb von τ_{gut}, wird der Punkt als schlecht deklariert. In Algorithmus 6 sind die Schritte der Qualitätsprüfung noch einmal zusammengefasst. Algorithmus 7 gibt einen Überblick über den gesamten Prozess der Merkmalspunktverfolgung.

Die Rechenzeit für den Überwachungsprozess hat durch die Reduzierung der Rechenoperationen für die affinen Transformationsparameter stark abgenommen. Ohne die Reduzierung dieser Rechenoperationen nahm der Überwachungsprozess in etwa das 4-fache der Rechenzeit des Verfolgungsprozesses ein. Abhängig von Faktoren wie Anzahl der zu verfolgenden Merkmalspunkte, Komprimierungsqualität der Videosequenz oder Struktur der 3D-Szene, war die Berechnung der Deformationsmatrix nur bei etwa 19% der Merkmalspunkte notwendig.

5.2.6. Ergebnisse

In diesem Abschnitt sollen nun die Auswirkungen der einzelnen Fehlerfunktionen und der Schwellwerte untersucht werden. Um eine Aussage treffen zu können, welche Fehlerfunktion sich mit welchem Schwellwert am besten eignet, um die Qualität der verfolgten Merkmalspunkte zu beurteilen, soll zunächst untersucht werden, wie sich die einzelnen

Ziel: Für alle Merkmalspunkte $\mathbf{x}_i$ im Bild I ihre entsprechende Position $\mathbf{x}'_i$ im nachfolgenden Bild J finden.

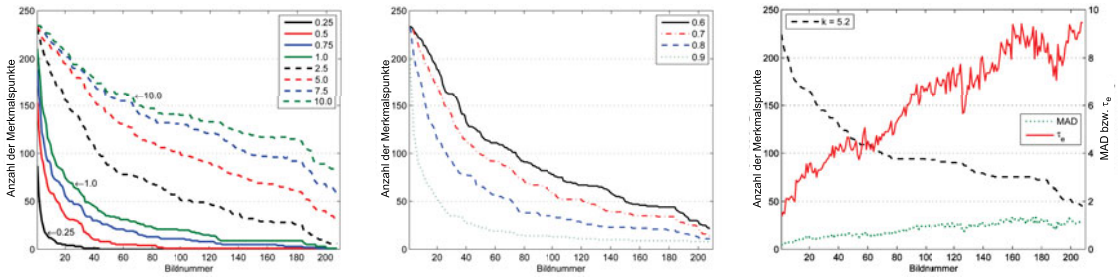
Wiederhole für alle Merkmalspunkte i

1. Bestimme die Position $\mathbf{x}'_i$ des Punktes $\mathbf{x}_i$ im nachfolgenden Bild J mittels Algorithmus 5.
2. Bestimme den Fehler für diesen Punkt $\mathbf{x}'_i$ zwischen dem Bild J und dem Bild I_n mittels Algorithmus 6.
3. Deklariere den Merkmalspunkt als „gut“ oder „schlecht“ entsprechend Algorithmus 6.

Ende der Schleife

Ergebnis: Die Positionen $\mathbf{x}'_i$ der Merkmalspunkte, welche als „gut“ deklariert wurden, im Bild J.

Algorithmus 7: Verfolgung und Überwachung von Merkmalspunkten in einer Videosequenz



(a) SSD (b) NCC (c) *X84*

Abbildung 5.5.: Anzahl der Merkmalspunkte über die Sequenz bei verschiedenen Fehlerfunktionen und Schwellwerten. In (c) wurde der Schwellwert für die Aussortierung von Merkmalspunkten nach der *X84-Rejection-Rule* bestimmt. Die gepunktete grüne Line zeigt den ermittelten MAD-Wert an, während die durchgezogene rote Line den resultierenden Schwellwert τ_e darstellt (Quelle: [193]).

Funktionen und Schwellwerte auf die Anzahl der in jedem Bild vorhandenen Merkmalspunkte auswirken. In Abbildung 5.5 wird dargestellt, wie viele Punkte in jedem Bild derselben Sequenz (*Palast*, 208 Bilder) bei unterschiedlichen Fehlerfunktionen und verschiedenen Schwellwerten pro Bild vorhanden sind. Bei der Bestimmung des Fehlers für einen Merkmalspunkt wurde die affine Transformation zum Ausgangsbild mitberücksichtigt.

Vor allem bei der Verwendung der SSD-basierten Fehlerfunktion ist es schwierig, einen geeigneten Schwellwert zu definieren, da der Fehler u.a. im hohen Maße von der gewählten Fenstergröße abhängt. Aus diesem Grund ist in Abbildung 5.5(c) der Verlauf dieses Schwellwertes durch Berechnung des MAD-Wertes ergänzend dargestellt. Die *X84-*

Rejection-Rule von Fusiello et al. [53] sollte vor allem bei SSD-basierten Fehlerfunktionen Verwendung finden.

In Abbildung 5.5(a) ist zu erkennen, dass, wenn der Schwellwert zu gering gewählt wird, schon nach wenigen Bildern kaum noch Merkmalspunkte vorhanden sind. Wird die Schwelle hingegen zu groß gewählt, werden keine oder kaum noch Punkte entfernt. Die Abnahme der Anzahl im Verlauf der Sequenz hängt dann fast ausschließlich damit zusammen, dass immer mehr Punkte aus den Bereich des Bildes wandern.

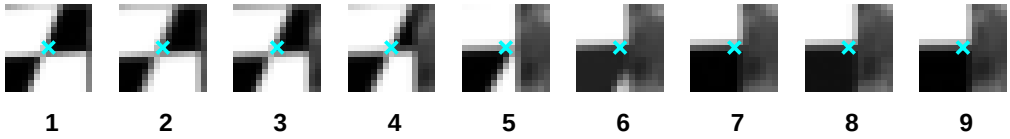
Wird die NCC-basierte Fehlerfunktion verwendet, ist eine Festlegung des Schwellwertes sehr viel einfacher. Allerdings kann auch hier eine schlechte Wahl des Schwellwertes dazu führen, dass zu viele oder zu wenige Merkmalspunkte entfernt werden. Dies wird in Abbildung 5.5(b) verdeutlicht. In weiteren Simulationen hat sich durch Beobachtung ein Schwellwert zwischen 0,7 und 0,8 als am Geeignetsten erwiesen.

Mithilfe der Fehlerfunktionen sollen vor allem Verdeckungen von Merkmalspunkten erkannt werden. In Abbildung 5.6(a) ist ein Beispiel für die Verdeckung eines Merkmalspunktes durch ein Vordergrundobjekt dargestellt. Der kritische Zeitpunkt, bei dem die Verdeckung auftritt, ist bei dem Übergang von Bild 4 zu Bild 5. Dies ist auch an den Verläufen der NCC- bzw. SSD-Werten in Abbildungen 5.6(b) und 5.6(c) erkennbar. Beim Übergang vom 4. zum 5. Bild haben beide Kurven betragsmäßig die größte Steigung. Ein NCC-Schwellwert von unter 0,6 würde z. B. erst ab dem 6. Bild den Merkmalspunkt entfernen, obwohl die genaue Position bereits im 5. Bild nicht mehr korrekt war.

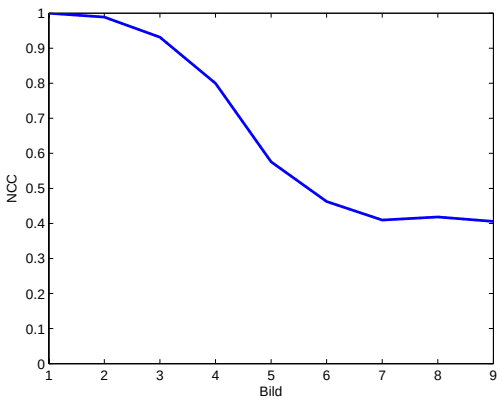
Für die nachfolgenden Untersuchungen wurde nur noch die NCC als Fehlerfunktion verwendet. Da die Berechnung der affinen Transformation für jeden Merkmalspunkt relativ viel Zeit im Vergleich zur Rechenzeit für die Verfolgung der Punkte in Anspruch nimmt, wurde untersucht, inwieweit sich ein Teil der Rechenzeit einsparen lässt. In Abschnitt 5.2.5 wurde vorgeschlagen, dass man die affine Transformation nur dann bestimmt, wenn sich der Wert der Fehlerfunktion zwischen einem oberen und einem unteren Schwellwert befindet. Bei der NCC gibt der obere Schwellwert an, welche Punkte als hinreichend gut klassifiziert werden können, ohne dass die affine Transformation bestimmt wird. Der untere Schwellwert hingegen bestimmt, ab wann Merkmalspunkte als zu schlecht klassifiziert werden, um für sie eine affine Transformation zu berechnen. Der Schwellwert τ_{gut} hat dabei eine direkte Auswirkung darauf, für wie viele Merkmalspunkte die affine Transformation bestimmt wird.

Es sollen nun 200 Merkmalspunkte über die Sequenz verfolgt werden, wobei in jedem Bild, in dem Punkte verloren gehen, neue Merkmalspunkte detektiert werden, damit die Anzahl an Punkten in jedem Bild nahezu konstant bleibt.

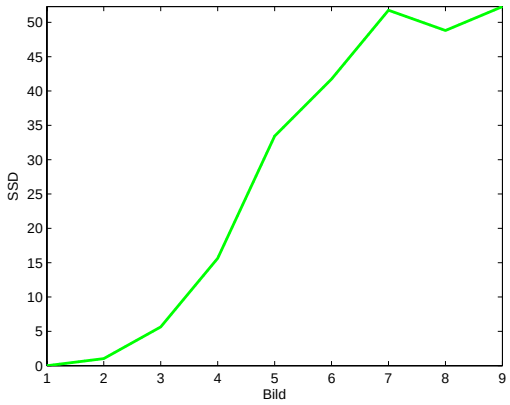
Aus der Tabelle 5.1 wird ersichtlich, dass ein zu hoher Wert für τ_{gut} dazu führt, dass ein Merkmalspunkt sehr schnell aussortiert wird. Bei einem Wert von 0,95 für τ_{gut} wird



(a)



(b) NCC



(c) SSD

Abbildung 5.6.: Zunehmende Verdeckung eines Merkmalspunktes (a) und die daraus resultierenden Werte der NCC (b) sowie der SSD (c) über das gesamte Fenster bezüglich dem ersten Bild (Quelle: [193]).

Tabelle 5.1.: Auswirkung des Schwellwertes τ_{gut} als obere Grenze, unterhalb der die affine Transformation berechnet wurde.

Schwellwert τ_{gut}	ges. Anzahl der Punkte über die Sequenz	durchschn. Anzeigedauer eines Punktes in Bildern	durchschn. pro Bild hinzugefügte Punkte	gesamte Rechenzeit normiert auf $\tau_{gut} = 0,5$
0,50	679	62,32	2,35	1
0,55	722	58,31	2,60	1,12
0,60	782	53,77	2,90	1,27
0,65	850	49,27	3,25	1,30
0,70	1000	42,08	3,98	1,47
0,75	1177	35,60	4,97	1,55
0,80	1471	28,41	6,52	1,66
0,85	1951	21,37	9,27	1,79
0,90	3081	13,31	17,81	1,99
0,95	5166	6,92	53,59	3,07

ein Punkt im Mittel nur für 6,92 Bilder als gut eingestuft. Später in diesem Kapitel wird gezeigt, dass diese Zahl zu gering ist, da der Abstand von etwa sieben Bildern noch zu gering für eine 3D-Rekonstruktion ist. Die Gesamtanzahl an detektierten und verfolgten Merkmalspunkten stieg bei diesem Wert auf über 5000 in der Sequenz.

Tabelle 5.2.: Einfluss der Größe des Schwellwertes τ_{gut} bei der Transformationfehlerbestimmung. Die affine Transformation wurde bestimmt, wenn der Transformationsfehler ohne affine Transformation 70% des Schwellwertes von τ_{gut} nicht unterschritt. Der Fehler wurde für ein 21×21 Pixel großes Fenster bestimmt.

Schwellwert τ_{gut}	ohne Transformation NCC $\geq \tau_{gut}$	mit Transformation NCC $\geq \tau_{gut}$	mit Transformation NCC $< \tau_{gut}$	ohne Transformation NCC $< \tau_{schlecht}$	Rechenzeit normiert auf $\tau_{gut} = 0,55$
0,55	37125	4165	563	134	1
0,6	36272	4858	684	142	1,12
0,65	35578	5329	891	127	1,19
0,7	35050	5571	1143	117	1,26
0,75	34240	6001	1487	106	1,34
0,8	33843	5797	2055	79	1,39
0,85	33529	5129	2985	86	1,47
0,9	32164	4216	5184	73	1,74
0,95	24705	2507	13555	90	2,83

Ein Wert von 0,5 für τ_{gut} bewirkt hingegen, dass ein Merkmalspunkt im Mittel 62,3 Bilder als gut klassifiziert wird. Wie aber in Abbildung 5.6 dargestellt wurde, ist dieser Wert zu niedrig, um Verdeckungen gut erkennen zu können.

Die starke Zunahme der Rechenzeit für $\tau_{gut} \geq 0,9$ entsteht vor allem durch das erneute detektieren von Merkmalspunkten in jedem Bild, um die entfernten Punkte wieder zu kompensieren.

Die Tabellen 5.2 und 5.3 zeigen die Auswirkung von verschiedenen Werten für die Schwellwerte τ_{gut} und $\tau_{schlecht}$. In Spalte 2 ist jeweils die Anzahl der Merkmalspunkte angegeben, für die der NCC-Wert in allen Bildern der Sequenz größer als der obere Schwellwert war. Die Spalten 3 und 4 geben die Anzahl der Punkte an, für die eine affine Transformation bestimmt wurde, wodurch der NCC-Wert über die Schwelle τ_{gut} (Spalte 3) gelangte bzw. unter dieser Schwelle (Spalte 4) blieb.

Der Tabelle 5.2 ist zu entnehmen, dass ein höherer Schwellwert zunehmend die Berechnung der affinen Transformation und damit eine Erhöhung der Rechenzeit bewirkt. An dieser Stelle sei darauf hingewiesen, dass für das Korrelationsfenster eine Größe von 21×21 Pixeln verwendet wurde. Eine kleinere Fenstergröße (15×15) führte dazu, dass die affine Transformation für keinen Merkmalspunkt den NCC-Wert über die Schwelle τ_{gut} gebracht hat. Die affinen Transformationsparameter ließen sich somit nicht bzw. nur geringfügig bestimmen. Das Fenster, welches für die Bestimmung der affinen Transformation verwendet wird, sollte demnach, wie in Abschnitt 5.2.2 erläutert, nicht zu klein gewählt werden, da sonst von vornherein auf die Bestimmung der affinen Transformationsparameter verzichtet werden kann.

Tabelle 5.3.: Auswirkung der unteren Schwelle $\tau_{schlecht}$ auf die Fehlerbestimmung. Als untere Schwelle, ab der eine affine Transformation berechnet wird, wurde $\tau_{schlecht} = \tau_{gut} \cdot p$ (mit $\tau_{gut} = 0,8$) verwendet. Der Fehler wurde für ein 21×21 Pixel großes Fenster bestimmt.

Prozentsatz p in %	ohne Transformation NCC $\geq \tau_{gut}$	mit Transformation NCC $\geq \tau_{gut}$	mit Transformation NCC $< \tau_{gut}$	ohne Transformation NCC $< \tau_{schlecht} = p \cdot \tau_{gut}$	Rechenzeit normiert auf $p = 95\%$
40	33577	6091	2100	6	1,36
45	33593	6074	2099	8	1,36
50	33724	5941	2096	13	1,33
55	33697	5962	2091	22	1,36
60	33699	5956	2079	39	1,35
65	33759	5885	2076	54	1,35
70	33843	5797	2055	79	1,34
75	34039	5600	2011	127	1,33
80	34377	5240	1958	200	1,30
85	34669	4920	1854	332	1,26
90	35432	4077	1661	601	1,18
95	36889	2541	1206	1134	1

Bei den weiteren Untersuchungen wurde der untere Schwellwert $\tau_{schlecht}$ prozentual abhängig von dem Schwellwert $\tau_{gut} = 0,8$ festgelegt. Die Ergebnisse sind in der Tabelle 5.3 dargestellt. Für einen prozentualen Wert von weniger als 70% des Wertes von τ_{gut} nimmt die Anzahl an Merkmalspunkten, welche unter Einbeziehung der affinen Transformation einen NCC-Wert erreichen, der über τ_{gut} liegt, nur noch geringfügig zu. Ein Wert von $p > 85\%$ führt dazu, dass verhältnismäßig viele Merkmalspunkte bereits ohne Transformations-Bestimmung als schlecht klassifiziert werden. Der Einfluss des unteren Schwellwerts auf die Rechenzeit ist kaum bemerkenswert. Für den interessierten Leser wird an dieser Stelle auf die Diplomarbeit von Walter [193] verwiesen, in der weiterführende Untersuchungen dargestellt wurden.

5.3. Detektieren von Schlüsselbildern

In diesem Abschnitt soll wieder die Problematik der 3D-Rekonstruktion im Vordergrund stehen. Nachdem Merkmalspunkte über den Verlauf der Videosequenz verfolgt wurden, gilt es nun den Kamerapfad und die 3D-Koordinaten der verfolgten Punkte zu bestimmen. Wie bereits in Abschnitt 5.1 erwähnt, ist eine 3D-Rekonstruktion aus benachbarten Bildern der Sequenz aufgrund des geringen Abstands der Kamerazentren nicht möglich. Ziel ist

es nun, Schlüsselbilder (engl. *keyframes*) zu detektieren, die einen hinreichend großen Abstand der Kamerazentren aufweisen, so dass Triangulationsverfahren angewendet werden können.

In Kapitel 2 wurden die theoretischen Grundlagen der 2D-Transformationen und der Epipolar-Geometrie vorgestellt. Dabei wurde gezeigt, dass eine 2D-Transformation zwischen zwei Bildern möglich ist, wenn

- die Kamera um das Kamerazentrum rotiert,
- die Kamera nur eine geringfügige translatorische Bewegung durchführt,
- die aufgenommene Szene einer Ebene im Raum entspricht.

Der letzte Fall kann insofern ausgeschlossen werden, da eine 3D-Rekonstruktion einer Ebene nicht möglich und als Anwendungszenario nicht von Interesse ist. Eine rein rotatorische Bewegung der Kamera über die gesamte Sequenz muss ebenfalls ausgeschlossen werden, da auch in diesem Fall eine 3D-Rekonstruktion nicht möglich wäre. Die Strahlen der 3D-Raumpunkte würden dabei immer durch dasselbe Kamerazentrum führen, wie in Abbildung 5.1 dargestellt wurde. Somit lassen sich die Strahlensätze zur Bestimmung der 3D-Koordinaten nicht anwenden. Geht man also davon aus, dass sich die Kamera zwischen benachbarten Bildern einer Videosequenz geringfügig translatorisch bzw. nur kurzzeitig rotatorisch bewegt hat, so lassen sich die verfolgten Merkmalspunkte durch ein zweidimensionales Bewegungsmodell (eine 2D-Transformation) beschreiben.

Im Verlauf der Sequenz nimmt allerdings der Abstand der Kamerazentren weiter zu. Damit steigt auch der Transformationsfehler bei Verwendung eines zweidimensionalen Bewegungsmodells. Die Bewegung der Kamera relativ zur statischen Szene muss dann durch ein dreidimensionales Bewegungsmodell beschrieben werden, welches durch die Epipolar-Geometrie repräsentiert wird. Torr [181] stellte dazu das *geometrisch robuste Informationskriterium* (engl. *geometric robust information criterion*, GRIC) als ein robustes Verfahren zur Auswahl des geeigneten Bewegungsmodells vor. Mit diesem Verfahren lassen sich Schlüsselbilder an der Stelle des Übergangs vom zweidimensionalen- zum dreidimensionalen Bewegungsmodell detektieren.

Ausgehend von den verfolgten Merkmalspunkten und dem ersten Bild der Sequenz, welches als erstes Schlüsselbild initialisiert wird, werden sowohl die Homographie H als auch die Fundamentalmatrix F zwischen dem aktuellen Bild der Sequenz und dem zuletzt detektierten Schlüsselbild (zu Beginn das erste Bild der Sequenz) ermittelt und der Transformationsfehler bezüglich einer jeweiligen Fehlerfunktion berechnet und verglichen. Das nächste Schlüsselbild wird dann detektiert, wenn der Transformationsfehler, der im Folgenden als GRIC-Wert bezeichnet wird, des 3D-Bewegungsmodells ($GRIC_F$) kleiner als der GRIC-Wert des 2D-Bewegungsmodells ($GRIC_H$) wird.

Die Fehlerfunktion ist allgemein definiert mit

$$GRIC = \sum_{i=1}^{n} \rho\left(e_i^2\right) + \lambda_1 Dn + \lambda_2 k, \tag{5.37}$$

wobei $\rho\left(e_i^2\right) = \min\left(\frac{e_i^2}{\sigma^2}, \lambda_3(r-d)\right)$ ist. Die Parameter sind wie folgt definiert: D ist die Dimension des gewählten Bewegungsmodells (H hat die Dimension zwei und F die Dimension drei), r ist die Dimension des Datensatzes (d.h. vier für zwei Ansichten), k ist die Anzahl der zu ermittelnden Modellparameter (sieben für F und acht für H), n ist die Anzahl der Punktkorrespondenzen zwischen den beiden Ansichten und σ ist die Standardabweichung der Transformationsfehler für alle Punktkorrespondenzen. e_i stellt den Transformationsfehler eines Merkmalspunktes zwischen den beiden Ansichten dar und wird je nach Bewegungsmodell berechnet mit

$$e_i = e_i^H = d(\mathbf{x}_i', H\mathbf{x}_i) \tag{5.38}$$

bzw.

$$e_i = e_i^F = d(\mathbf{x}_i', F\mathbf{x}_i). \tag{5.39}$$

$d(\mathbf{x}_i', H\mathbf{x}_i)$ entspricht dabei dem Euklidischen Abstand zwischen dem Punkt $\mathbf{x}_i$ im Schlüsselbild, der über eine planare Homographie in die aktuelle Ansicht transformiert wurde, und dem Punkt $\mathbf{x}_i'$ im aktuellen Bild. Für $d(\mathbf{x}_i', F\mathbf{x}_i)$ gilt entsprechendes, nur dass $F\mathbf{x}_i = \mathbf{l}_i'$ die Epipolarlinie im aktuellen Bild ist und somit der Euklidische Abstand des Punktes $\mathbf{x}_i'$ zur Epipolarlinie $\mathbf{l}_i'$ berechnet wird. Die Parameter λ_1, λ_2 und λ_3 sind Regelungsparameter und werden in [185] mit $\lambda_1 = \log 4$, $\lambda_2 = \log 4n$ und $\lambda_3 = 2$ angegeben.

An dieser Stelle sollte darauf hingewiesen werden, dass e_i^F immer kleinere Werte annimmt als e_i^H, da die Abbildung von Punkten auf Linien, wie bei der Epipolar-Geometrie, die allgemein gültigere Transformation ist als die Abbildung von Punkten auf Punkte. Der Term $\lambda_1 dn$ in Gl. (5.37) soll dieses aber genau kompensieren und stellt für die Berechnung des GRIC-Wertes eine Art „Bestrafung“ des Bewegungsmodells höherer Ordnung dar. Wie bereits in Kapitel 2 gezeigt wurde, sind für die Berechnung der Matrix H mindestens vier Punktkorrespondenzen und für die Berechnung von F mindestens sieben bzw. acht Punktkorrespondenzen notwendig.

In Abbildung 8.4 ist ein Beispiel für die Detektion von Schlüsselbildern aus 200 Einzelbildern der TV-Sequenz *Palast* dargestellt. Die statische Szene wurde mit einer Kamera aus einem Hubschrauber aufgenommen. Insgesamt wurden 13 Schlüsselbilder detektiert. Der Verlauf der Differenzen der beiden GRIC-Werte ($GRIC_F - GRIC_H$) ist ebenfalls in dieser Abbildung dargestellt. Ein Schlüsselbild wurde dann detektiert, wenn die Differenz einen negativen Wert angenommen hat.

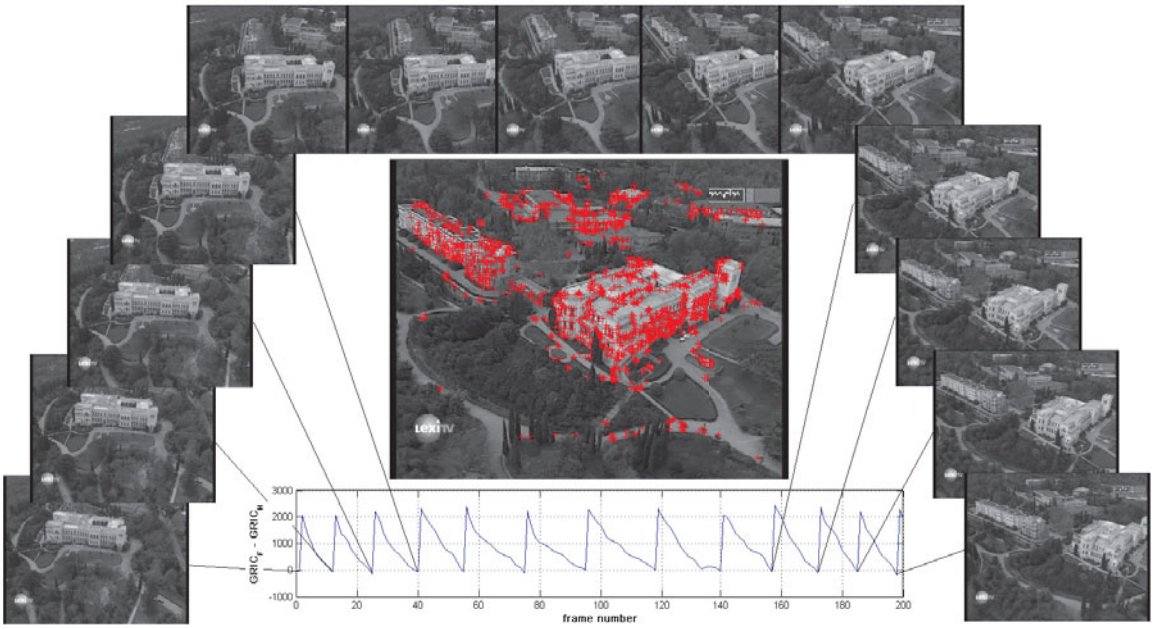


Abbildung 5.7.: Detektion von 13 Schlüsselbildern aus der *Palast*-Sequenz mit insgesamt 200 Bildern und Darstellung des Verlaufs der GRIC-Werte.

Erweiterung des GRIC-Ansatzes. In [141] stellen Pollefeys et al. eine Erweiterung des GRIC-Ansatzes von Torr vor, in der nicht sofort ein Schlüsselbild detektiert wird, wenn die Differenz $GRIC_F - GRIC_H$ negativ wird, da zu diesem Zeitpunkt ja gerade der Übergang der beiden Bewegungsmodelle stattfindet. In Abbildung 4.1 wurde gezeigt, dass der Toleranzbereich für die Triangulation der 3D-Raumpunkte mit zunehmendem Abstand der Kamerazentren kleiner wird. Eine spätere Detektion der Schlüsselbilder scheint demnach also sinnvoll zu sein, um die Qualität der Rekonstruktionsergebnisse zu verbessern. Aus dieser Sicht stellt sich allerdings die Frage, warum nicht gleich die beiden am weitesten entfernten Kamerapositionen verwendet werden? Im Falle einer translatorischen Bewegung der Kamera in nur eine Richtung wären das das erste und das letzte Bild der Sequenz. Die Frage ist schnell beantwortet, denn es ist nicht davon auszugehen, dass Merkmalspunkte, die im ersten Bild der Sequenz detektiert wurden, über die gesamte Sequenz verfolgt werden können. Im schlimmsten Fall existiert ggf. keiner dieser Punkte in der letzten Ansicht.

Ein zweites Kriterium, welches die Qualität der Rekonstruktionsergebnisse verbessert, ist die Anzahl der zur Verfügung stehenden Merkmalspunkte. Rauscheinflüsse können durch eine größere Anzahl an Punkten und mithilfe robuster Schätzverfahren (siehe Kapitel 3) minimiert werden.

Das Verfahren von Pollefeys et al. [141] berücksichtigt beide Kriterien, indem ein Schlüsselbild detektiert wird, wenn die Anzahl der verfolgten Merkmalspunkte noch 90% der Anzahl der Merkmalspunkte entspricht, die zum Zeitpunkt des Nulldurchgangs

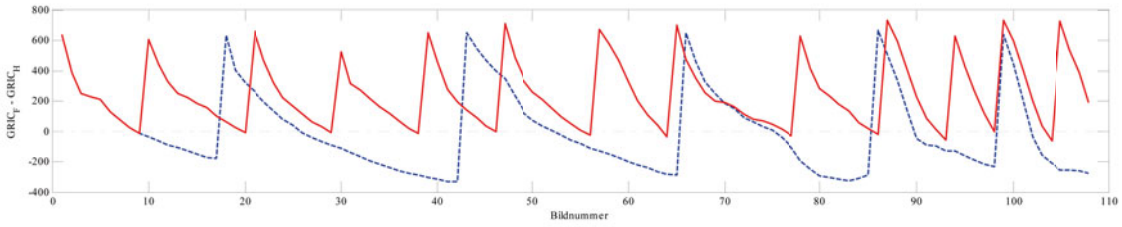


Abbildung 5.8.: GRIC-Verläufe der TV-Sequenz *Felsen* mit 109 Bildern. Rot: $GRIC_F - GRIC_H$ nach Torr et al. [185], blau: $GRIC_F - GRIC_H$ und 90% -Schwelle nach Pollefeys et al. [141]

($GRIC_F - GRIC_H$) verfolgt wurden. In Abbildung 5.8 werden die beiden Verfahren nach Torr et al. [185] und Pollefeys et al. [141] anhand der GRIC-Verläufe der TV-Sequenz *Felsen* mit einer Gesamtlänge von 109 Bildern gegenübergestellt. Nach Torr et al. sind insgesamt 13 Schlüsselbilder (inkl. erstes Bild der Sequenz) detektiert worden (rote Kurve). Bei dem Verfahren von Pollefeys et al. (blaue Kurve) wurden lediglich sechs Schlüsselbilder (inkl. erstes Bild der Sequenz) detektiert. Der Abstand zwischen den Schlüsselbildern hat sich dabei in etwa verdoppelt.

Eine andere Erweiterung des GRIC-Ansatzes wurde in [177] vorgeschlagen, bei der anstelle des 90%-Schwellwerts nach Pollefeys et al. eine initiale Schlüsselbildpaarung über die Reprojektionsfehler, die für jede mögliche Schlüsselbildpaarung ($GRIC_F < GRIC_H$) ermittelt werden, gefunden wird.

5.4. Initiale Rekonstruktion

In Kapitel 4 wurde beschrieben, wie aus den Punktkorrespondenzen zwischen zwei Ansichten die 3D-Struktur bestimmt werden kann. Videosequenzen enthalten dagegen zusätzliche Ansichten, die entweder zwischen diesen Ansichten liegen oder auf diese folgen. Mithilfe der zusätzlichen Informationen kann somit sowohl die 3D-Rekonstruktion verbessert als auch die 3D-Struktur um weitere 3D-Punkte erweitert werden. Beardsley et al. [9] sowie Pollefeys et al. [142] verwenden einen sequentiellen Algorithmus, um eine projektive Rekonstruktion der Szene zu ermitteln und schrittweise zu erweitern. Ausgangspunkt dabei ist jeweils eine initiale Rekonstruktion aus zwei Ansichten.

Zu Beginn wird eine initiale 3D-Rekonstruktion mit den Verfahren, wie sie in Kapitel 4 beschrieben wurden, durchgeführt. Dazu werden die ersten beiden Schlüsselbilder verwendet, die, wie in Abschnitt 5.3 beschrieben wurde, mithilfe des GRIC-Ansatzes detektiert wurden.

Aus den Punktkorrespondenzen wird zunächst die Fundamentalmatrix mit einem robusten Ansatz ermittelt und die Projektionsmatrizen P_1 und P_2 berechnet, wobei das Weltkoordinatensystem hier in das Kamerazentrum der ersten Kamera gelegt wird. Damit ist P_1:

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Die 3D-Koordinaten der Punkte werden anschließend mit der optimalen Triangulation (siehe Abschnitt 4.6) bestimmt.

5.5. Hinzufügen weiterer Ansichten

Nachdem die initiale 3D-Struktur und die Projektionsmatrizen P_1 und P_2, wie im vorangegangenen Abschnitt kurz beschrieben, bestimmt wurden, sollen nun weitere Ansichten der Szene dem 3D-Modell hinzugefügt werden. Dazu werden vorerst die nächst folgenden Schlüsselbilder und anschließend die Zwischenansichten verwendet. Da das Hinzufügen neuer Ansichten und die damit verbundenen Rechenoperationen einen sehr komplexen Vorgang darstellen, sollen die Arbeitsschritte kurz zusammengefasst werden und einen Überblick für diesen Abschnitt liefern:

- Schrittweises Hinzufügen der detektierten Schlüsselbilder
- Bestimmung der Kameraprojektionsmatrizen über 2D-3D-Punktkorrespondenzen
- Bestimmung weiterer 3D-Raumpunkte mittels Triangulation
- Optimierung der Kameraparameter und der 3D-Koordinaten der Raumpunkte

5.5.1. Bestimmung der Kameraprojektionsmatrizen über 2D-3D-Punktkorrespondenzen

Um die Projektionsmatrix einer weiteren Ansicht bestimmen zu können, müssen in dieser die 2D-Koordinaten von einer genügend großen Anzahl von schon berechneten 3D-Punkten aus der Initialisierungsphase bekannt sein, d.h. Merkmalspunkte die über die Sequenz verfolgt wurden, müssen sowohl in dem aktuellen Schlüsselbild als auch in den beiden vorherigen Schlüsselbildern vorhanden sein. Die daraus resultierenden 2D-3D-Beziehungen lassen sich in homogenen Koordinaten angeben mit

$$\mathbf{x}_{ij} = P_j \mathbf{X}_i, \quad \text{mit } \mathbf{x}_{ij} = \begin{pmatrix} x_{ij} \\ y_{ij} \\ 1 \end{pmatrix} \text{ und } \mathbf{X}_i = \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix}, \tag{5.40}$$

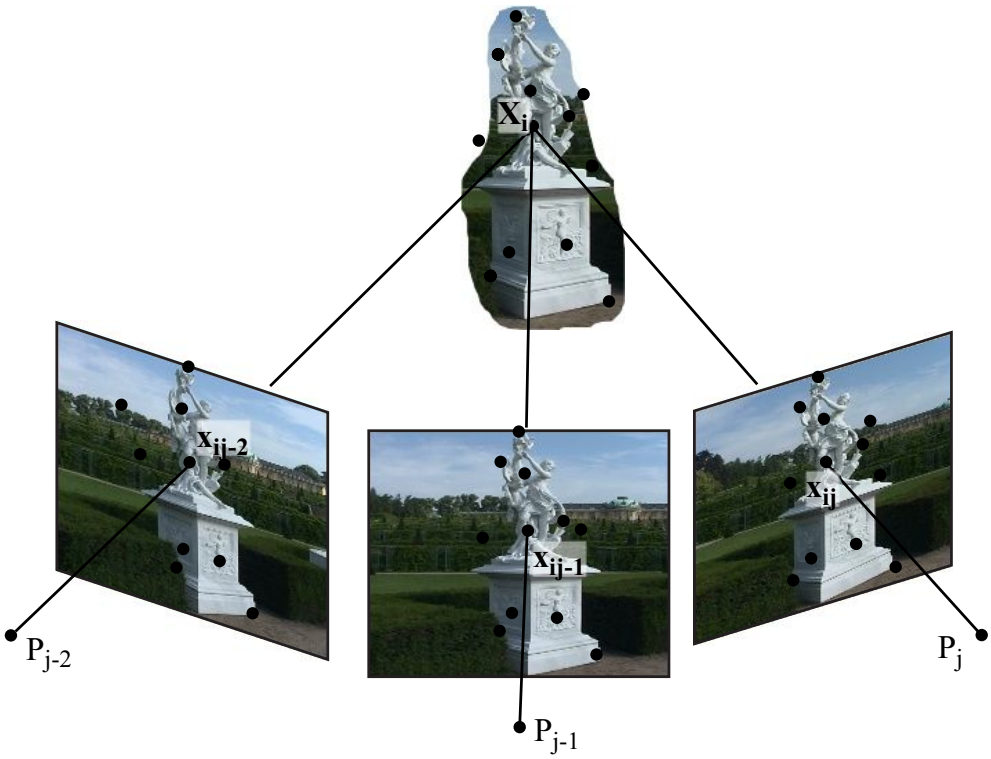


Abbildung 5.9.: Mit den aus den vorangegangenen Bildern $j-1$ und $j-2$ bestimmten 3D-Punkten und ihren bekannten Positionen im Bild j wird die Projektionsmatrix P_j berechnet.

wobei $\mathbf{x}_{ij}$ den i-ten Punkt in der j-ten Ansicht beschreibt. Abbildung 5.9 zeigt ein Beispiel für das Hinzufügen einer Ansicht j mit der Projektionsmatrix P_j. Gl. (5.40) kann wie folgt umformuliert werden:

$$\begin{aligned} x_{ij} &= \mathbf{p}^{1T}\mathbf{X}_i \\ y_{ij} &= \mathbf{p}^{2T}\mathbf{X}_i \quad \text{mit } \mathbf{p}^{kT} \text{ der k-ten Zeile der Matrix } P_j \\ 1 &= \mathbf{p}^{3T}\mathbf{X}_i \end{aligned} \tag{5.41}$$

Durch Umstellen der Gl. (5.41) erhält man das lineare Gleichungssystem

$$\begin{aligned} x_{ij}\mathbf{p}^{3T}\mathbf{X}_i - \mathbf{p}^{1T}\mathbf{X}_i = 0 \quad \text{und} \\ y_{ij}\mathbf{p}^{3T}\mathbf{X}_i - \mathbf{p}^{2T}\mathbf{X}_i = 0. \end{aligned}$$

bzw. in Matrix-Schreibweise

$$\begin{bmatrix} -\mathbf{X}_i^T & \mathbf{0}^T & x_{ij}\mathbf{X}_i^T \\ \mathbf{0}^T & -\mathbf{X}_i^T & y_{ij}\mathbf{X}_i^T \end{bmatrix} \begin{pmatrix} \mathbf{p}^1 \\ \mathbf{p}^2 \\ \mathbf{p}^3 \end{pmatrix} = \mathbf{0}. \tag{5.42}$$

Aus n gegebenen 2D-3D-Korrespondenzen lässt sich ein lineares Gleichungssystem mit $2n$ Gleichungen aufstellen:

$$\begin{bmatrix} -\mathbf{X}_1^T & \mathbf{0}^T & x_{1j}\mathbf{X}_1^T \\ \mathbf{0}^T & -\mathbf{X}_1^T & y_{1j}\mathbf{X}_1^T \\ -\mathbf{X}_2^T & \mathbf{0}^T & x_{2j}\mathbf{X}_2^T \\ \mathbf{0}^T & -\mathbf{X}_2^T & y_{2j}\mathbf{X}_2^T \\ \vdots & \vdots & \vdots \\ -\mathbf{X}_n^T & \mathbf{0}^T & x_{nj}\mathbf{X}_n^T \\ \mathbf{0}^T & -\mathbf{X}_n^T & y_{nj}\mathbf{X}_n^T \end{bmatrix} \begin{pmatrix} \mathbf{p}^1 \\ \mathbf{p}^2 \\ \mathbf{p}^3 \end{pmatrix} = A\mathbf{p} = \mathbf{0}. \tag{5.43}$$

Die Matrix A hat somit $2n \times 12$ Einträge. Der Vektor $\mathbf{p} = \begin{pmatrix} \mathbf{p}^1 & \mathbf{p}^2 & \mathbf{p}^3 \end{pmatrix}$ enthält alle zwölf Elemente der Projektionsmatrix P_j und hat elf Freiheitsgrade. Daraus ergibt sich, dass mindestens sechs 2D-3D-Punktkorrespondenzen mit jeweils zwei Gleichungen zur Lösung des Gleichungssystems nötig sind. Aufgrund von Rauscheinflüssen sollten allerdings mehr Korrespondenzen zur Bestimmung der Matrix P_j herangezogen werden. Analog zur Bestimmung der Fundamentalmatrix in Abschnitt 3.1.1 lässt sich ein DLT-Algorithmus anwenden, der den algebraischen Fehler und damit das Gleichungssystem $\|A\mathbf{p}\|$ unter der Bedingung $\|\mathbf{p}\| = 1$ minimiert.

Für die Bestimmung der Matrix P_j sollten die Daten wie bei dem normierten 8-Punkt-Algorithmus (Algorithmus 2) normiert werden. Zuerst werden alle 2D-Punkte über die Transformationsmatrix T normiert, so dass sich ihr Schwerpunkt im Koordinatenursprung befindet und der Abstand zum Koordinatenursprung $\sqrt{2}$ entspricht. Eine ähnliche Normalisierung wird auch für die 3D-Punkte über die Transformationsmatrix T' vorgenommen, nur dass ihr Abstand zum Koordinatenursprung $\sqrt{3}$ beträgt. Algorithmus 8 fasst die Bestimmung von P_j mittels normiertem DLT-Algorithmus zusammen.

Aus den selben Gründen, wie bei der Schätzung der Fundamentalmatrix, sollte allerdings zur Bestimmung der Projektionsmatrix ein robustes Verfahren verwendet werden, welches sich nicht ausschließlich auf eine Singulärwertszerlegung beschränkt. Hierzu wird erneut ein RANSAC-basierter Ansatz gewählt, wie er in Algorithmus 4 für die Bestimmung von F zum Einsatz kam. Ziel des robusten Ansatzes ist es wiederum, Ausreißer zu eleminieren, die die Berechnung von P_j negativ beeinflussen. Um entscheiden zu können, ob es sich bei einem Punkt um einen Ausreißer handelt, wird der Reprojektionsfehler des Punktes $\mathbf{X}_i$ in der Ansicht j als Fehlermaß verwendet. Für den Reprojektionsfehler gilt:

$$e_i = d(\mathbf{x}_{ij}, P_j\mathbf{X}_i)^2, \tag{5.44}$$

mit $d(\mathbf{x}_{ij}, P_j\mathbf{X}_i)$ dem Abstand zwischen dem 2D-Punkt $\mathbf{x}_{ij}$ und der Projektion des korrespondierenden 3D-Punktes $\mathbf{X}_i$ in die Bildebene unter Verwendung der Projektionsmatrix P_j.

Ziel: Bestimmung der Projektionsmatrix P_j für eine gegebene Anzahl $n > 6$ von 2D-3D-Punktkorrespondenzen $\{\mathbf{x}_{ij} \leftrightarrow \mathbf{X}_i\}$, so dass die Gleichung $\mathbf{x}_{ij} = P_j\mathbf{X}_i$ erfüllt wird.

1. **Normalisierung**: Transformiere die Bildkoordinaten nach $\hat{\mathbf{x}}_{ij} = T\mathbf{x}_{ij}$ und $\hat{\mathbf{X}}_i = T'\mathbf{X}_i$, wobei T und T' normalisierende Transformationen mit einer Verschiebung und Skalierung sind, so dass der Schwerpunkt aller Punkte im Koordinatenursprung $(0,0)^T$ liegt und der quadratische Mittelwert der Entfernung aller 2D-Punkte zum Koordinatenursprung $\sqrt{2}$ und der aller 3D-Punkte zum Koordinatenursprung $\sqrt{3}$ beträgt.
2. Bestimme für jede 2D-3D-Punktkorrespondenz $\hat{\mathbf{x}}_{ij} \leftrightarrow \hat{\mathbf{X}}'_i$ die Matrix $\hat{A}_i$ mittels Gleichung 5.42.
3. Setze alle n 2×12 Matrizen $\hat{A}_i$ zu einer $2n \times 12$ Matrix $\hat{A}$ zusammen.
4. Führe eine SVD von $\hat{A}$ durch. Der Vektor, der dem kleinsten Singulärwert entspricht, ist die Lösung von $\hat{\mathbf{p}}$. Ist $\hat{A} = UDV^T$, wobei D eine Diagonalmatrix mit positiven – in absteigender Reihenfolge – sortierten Werten auf der Diagonalen ist, entspricht $\hat{\mathbf{p}}$ der letzten Spalte von V.
5. Setze die Projektionsmatrix $\hat{P}_j$ aus dem Vektor $\hat{\mathbf{p}}$ zusammen.
6. **Denormalisierung**: Setze $P_j = T'^{-1}\tilde{P}'_j T$. Die resultirende Matrix P_j ist die Projektionsmatrix, die zu den Ausgangsdaten $\{\mathbf{x}_i \leftrightarrow \mathbf{X}'_i\}$ gehört.

Ergebnis: Die Projektionsmatrix P_j, für die der Vektor $\mathbf{p}$ das Gleichungssystem $\|A\mathbf{p}\|$ unter der Bedingung $\|\mathbf{p}\| = 1$ minimiert.

Algorithmus 8: Normierter DLT-Algorithmus für die Bestimmung der Projektionsmatrix P

Ziel: Robuste Lösung für die Projektionsmatrix P_j, wobei die Anzahl an gegebenen 2D-3D-Punktkorrespondenzen $\{\mathbf{x}_{ij} \leftrightarrow \mathbf{X}_i\}$ $n > 6$ ist.

1. **Wiederhole für** $k = 1$ **bis** N, wobei N mittels Gl. (3.17) bestimmt wird.
   a) Wähle zufällig einen Satz von 6 Punktkorrespondenzen aus und bestimme die Projektionsmatrix P_k mit Algorithmus 8.
   b) Bestimme für jede Punktkorrespondenz ihren Reprojektionsfehler e_{ik} mit Gl. (5.44).
   c) Bestimme die Anzahl der Punkte, deren Fehler e_{ik} kleiner als der Grenzwert τ_e ist: $e_{ik} < \tau_e$.

   Ende der Schleife
2. Wähle die Matrix P_k aus, bei der die meisten Punkte (Inlier) innerhalb des Toleranzbereiches liegen. Wenn es mehr als eine Matrix gibt, wähle die Matrix P_k, bei der die Summe E_k der Reprojektionsfehler aller Inlier am kleinsten ist.
3. Bestimme P_j mit den ermittelten Inliern.

Ergebnis: Projektionsmatrix P_j sowie die dazugehörigen Inlier.

Algorithmus 9: RANSAC-Algorithmus für die Bestimmung der Projektionsmatrix P_j unter Verwendung des normierten DLT-Algorithmus

5.5.2. Bestimmung weiterer 3D-Raumpunkte mittels Triangulation

Nachdem die Projektionsmatrix für ein weiteres Schlüsselbild ermittelt wurde, können nun in einem nächsten Schritt zusätzliche 3D-Punkte, die z. B. während des Verfolgungsprozesses erst im Verlauf der Sequenz detektiert wurden, mittels Triangulation bestimmt werden. Bei diesen Punkten handelt es sich um 2D-Punktkorrespondenzen zwischen dem letzten und aktuellen Schlüsselbild ($j-1$ und j), welche im Bild $j-2$ bzw. in weiter zurückliegenden Bildern noch nicht vorhanden waren.

Für diese 2D-Punktkorrespondenzen $\left\{\mathbf{x}_{ij-1} \leftrightarrow \mathbf{x}_{ij}\right\}$ können über die optimale Triangulation mit den beiden Projektionsmatrizen P_{j-1} und P_j, wie in Abschnitt 4.6 beschrieben, neue 3D-Punkte $\mathbf{X}_i$ bestimmt werden. Die dabei benötigte Fundamentalmatrix kann aus den beiden Projektionsmatrizen P_{j-1} und P_j mit Gl. (3.32) berechnet werden.

Da Ausreißer in den neuen Punktkorrespondenzen die Ergebnisse der Triangulation verschlechtern, sollten diese von vornherein identifiziert und von der Triangulation ausgenommen werden. Dazu kann erneut der RANSAC-Algorithmus für die Bestimmung der Fundamentalmatrix (Algorithmus 4) verwendet werden.

In [140] wird die 3D-Struktur noch einmal um zusätzliche 3D-Punkte erweitert, indem mithilfe der berechneten Fundamentalmatrix zwischen den beiden Ansichten $j-1$ und j und einer polaren Rektifizierung [143] nach neuen Punktkorrespondenzen gesucht wird (siehe Abschnitt 4.3.3).

5.5.3. Optimierung der Kameraparameter und der 3D-Koordinaten der Raumpunkte

Bei den vorherigen Betrachtungen wurden Projektionsmatrizen zusätzlicher Ansichten (Schlüsselbilder) über zuvor bestimmte 2D-3D-Punktkorrespondenzen ermittelt und weitere 3D-Punkte durch Triangulation erzeugt. Die Ergebnisse der Berechnungen der Projektionsmatrizen sind damit direkt bzw. die 3D-Punkte, die über diese Projektionsmatritzen berechnet wurden, indirekt von der initialen Rekonstruktion aus den ersten beiden Ansichten abhängig. Da die initiale Rekonstruktion nur auf einer Schätzung beruht, sind zumindest geringe Fehler nicht auszuschließen. Durch das Hinzufügen neuer Ansichten und die Bestimmung weiterer 3D-Punkte pflanzen sich diese Fehler weiter fort. Abbildung 5.10 zeigt dazu ein Beispiel einer Rekonstruktion aus mehreren Ansichten der Sequenz *Statue*. Die rote Kurve zeigt den Verlauf der Reprojektionsfehler in Abhängigkeit von der Bildnummer.

Für die Verbesserung der Rekonstruktionsergebnisse lassen sich eine Reihe von Optimierungsverfahren anwenden, die man in zwei Kategorien aufteilen kann: lineare Optimierung und nichtlineare Optimierung. Nichtlineare Optimierungverfahren sind aufgrund der hohen Anzahl der zu ermittelnden Parameter (zwölf für jede Projektionsmatrix

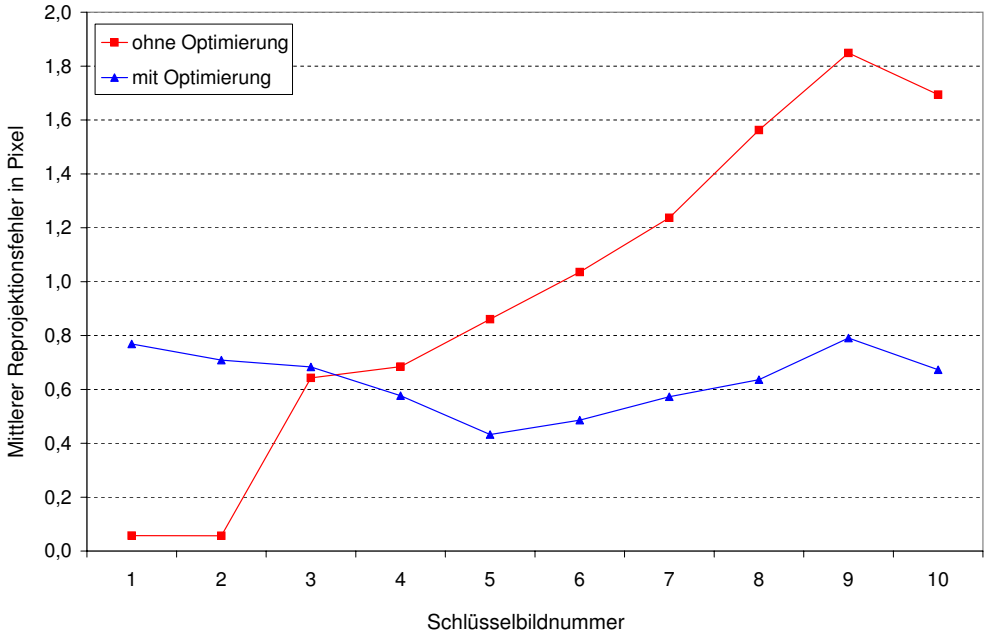


Abbildung 5.10.: Verlauf der mittleren Reprojektionsfehler von 248 (ohne Optimierung) bzw. 259 (mit Optimierung) ermittelten 3D-Punkten. Verwendet wurden 360 Bilder der Sequenz *Statue*, bei der 10 Schlüsselbilder detektiert wurden. Die Anzahl der verfolgten Punkte wurde auf 400 festgelegt. Der mittlere Reprojektionsfehler über alle Schlüsselbilder beträgt 0,9679 Pixel (ohne Optimierung) bzw. 0,6327 (mit Optimierung unter Verwendung des gleitenden Mittelwerts)

und drei für jeden 3D-Punkt) sehr rechenintensiv, liefern allerdings in der Regel die besten Rekonstruktionsergebnisse.

Lineares Optimierungsverfahren

Walter stellt in [193] ein einfaches lineares Optimierungsverfahren vor, welches auf dem gleitenden Mittelwert der Rekonstruktionsergebnisse beruht. Die Punktkorrespondenzen, mit denen die Projektionsmatrix P_j bestimmt wurde, können des Weiteren dazu verwendet werden, die bis dahin bestimmten Rekonstruktionsergebnisse zu verbessern. Hierzu werden für die 2D-2D-Korrespondenzen $\{\mathbf{x}_{ij-1} \leftrightarrow \mathbf{x}_{ij}\}$, die zwischen den beiden Bildern I_{j-1} und I_j vorhanden sind und für die bereits entsprechende 3D-Punkte $\mathbf{X}_i$ ermittelt wurden, unabhängig von den anderen Ansichten die zugehörigen 3D-Punkte $\mathbf{X}_i$ erneut durch Triangulation bestimmt. Anschließend wird für jeden dieser 3D-Punkte der gleitende Mittelwert mit den Ergebnissen der vorherigen Berechnungen wie folgt bestimmt:

$$\overline{\mathbf{X}}_i^j = \frac{(n-1)\overline{\mathbf{X}}_i^{j-1} + \mathbf{X}_i^j}{n}. \tag{5.45}$$

Dabei entspricht n der Anzahl an Ansichten, mit denen der Punkt $\mathbf{X}_i$ bis zu diesem Zeitpunkt bestimmt wurde. j ist der Index des aktuellen Bildes. Abbildung 5.10 zeigt den Verlauf des mittleren Reprojektionsfehlers bei Verwendung dieses linearen Optimierungsverfahrens (blaue Kurve) im Vergleich zu dem Reprojektionsfehler ohne Optimierung (rote Kurve). Zwar hat sich der Reprojektionsfehler in den ersten beiden Bildern vergrößert, dafür bleibt der Fehler für die weiteren Schlüsselbilder annährend konstant.

Nichtlineare Optimierungsverfahren

Nichtlineare Optimierungsverfahren sind, wie bereits erwähnt, äußerst rechenintensiv, können aber die Rekonstruktionsergebnisse noch entscheidend verbessern. Die Anzahl der zu optimierenden Parameter hängt dabei im Wesentlichen von der Länge der Videosequenz ab.

Beispiel: Betrachtet man eine Videosequenz mit 200 Bildern, so ergibt sich eine Anzahl an Parametern allein für die Projektionsmatrizen (zwölf Parameter pro Ansicht) von $N_a = 200 \cdot 12 = 2400$. Die Zahl der berechneten 3D-Punkte variiert sehr stark mit dem Texturgehalt des Datensatzes bzw. mit der Anzahl der verfolgten Merkmalspunkte. Werden z. B. 1000 Punkte rekonstruiert, so ergibt sich bei drei Parametern pro 3D-Punkt eine Anzahl $N_b = 1000 \cdot 3 = 3000$ und damit eine Gesamtzahl an zu optimierenden Parametern von $N_{gesamt} = N_a + N_b = 5400$.

Das gleichzeitige Optimieren der Parameter wird auch als *Bündelanpassung* (engl. *bundle adjustment*, BA) bezeichnet [188]. Beim Bundle Adjustment werden sowohl die 3D-Struktur als auch die Projektionsmatrizen so angepasst, dass der Reprojektionsfehler in allen Ansichten minimiert wird. Die zu minimierende Kostenfunktion ist gegeben mit:

$$E(P, \mathbf{X}) = \sum_{\forall i} \sum_{\forall j} d\left(\mathbf{x}_{ij}, P_j \mathbf{X}_i\right)^2 . \tag{5.46}$$

Die Funktion d ermittelt den Euklidischen Abstand zwischen dem 2D-Punkt $\mathbf{x}_{ij}$ und der Projektion des entsprechenden 3D-Punktes $\mathbf{X}_i$ in Bild I_j.

Für die nichtlineare Minimierung der Kostenfunktion können Verfahren wie z. B. das Gradientenabstiegs- oder das Gauß-Newton-Verfahren angewendet werden. Allgemein geläufig ist hierbei allerdings der Levenberg-Marquardt-Algorithmus (LM) [127], welcher eine Kombination aus dem Gradientenabstiegs- und dem Gauß-Newton-Verfahren darstellt. Die Berechnungsschritte des LM-Algorithmus sind im Anhang A.2 zusammengefasst.

In [115] wird ein Verfahren vorgestellt, welches den Rechenaufwand des Bundle-Adjustment reduziert, indem die Eigenschaft ausgenutzt wird, dass nicht jeder der berechneten 3D-Punkte in allen Ansichten einen korrespondierenden 2D-Punkt besitzt und

somit die verwendeten Matrizen in ihrer Größe reduziert werden können. Dieses Verfahren wird auch als *„spärliche" Bündelanpassung* (engl. *sparse-bundle-adjustment*, SBA) bezeichnet. Die Herleitung von SBA und die Berechnungsschritte sind in [115] ausführlich beschrieben.

5.5.4. Implementierungsvarianten

In diesem Abschnitt sollen nun verschiedene Implementierungsmöglichkeiten zur schrittweisen Rekonstruktion der 3D-Struktur und der Projektionsmatrizen vorgestellt und verglichen werden. Die unterschiedlichen Möglichkeiten beruhen dabei vor allem auf dem Einsatz linearer und nichtlinearer Optimierungsverfahren, welche für verschiedene Arbeitsprozesse variabel verwendet werden können. Ein Beispiel dafür ist, dass vorerst alle Rekonstruktionsschritte mit linearen Minimierungsverfahren operieren. Erst in einem abschließenden letzten Schritt wird eine globale Optimierung mit nichtlinearem Minimierungsverfahren durch Lösen der Gl. (5.46) durchgeführt. Die zu optimierende Anzahl der Parameter wäre dabei je nach Länge der Sequenz - wie im Beispiel in Abschnitt 5.5.3 gezeigt - recht groß. Mit zunehmender Komplexität des Minimierungsproblems steigt auch das Risiko, statt dem globalen Minimum, ein lokales Minimum zu treffen.

Lineare Rekonstruktionsmethode

Die erste Implementierungsmethode sieht eine rein lineare Rekonstruktion vor. Zunächst wird eine initiale Rekonstruktion der 3D-Struktur und der Projektionsmatrizen mithilfe der ersten beiden Schlüsselbilder - wie in Abschnitt 5.4 beschrieben - durchgeführt. Anschließend wird die Rekonstruktion bezüglich der 3D-Struktur und der Projektionsmatrizen für alle weiteren Schlüsselbilder erweitert (siehe Abschnitte 5.5.1 und 5.5.2). Die Bestimmung der Projektionsmatrizen erfolgt über das robuste Verfahren nach Algorithmus 9. Für die Optimierung der 3D-Struktur wird der gleitende Mittelwert der Rekonstruktionsergebnisse nach [193] in Abschnitt 5.5.3 verwendet. In einem letzten Schritt werden die Projektionsmatrizen der Ansichten zwischen den Schlüsselbildern über 2D-3D-Korrespondenzen mit Algorithmus 8 ermittelt.

Abbildung 5.11(a) fasst den Ablauf der Rechenschritte in einem Flussdiagramm zusammen.

Rekonstruktion mit nichtlinearer Optimierung der Projektionsmatrizen

Die zweite Implementierungsvariante sieht eine einfache Erweiterung der linearen Rekonstruktionsmethode vor, indem sämtliche Projektionsmatrizen, die über 2D-3D-Punktkorrespondenzen berechnet wurden, einzeln mithilfe des Levenberg-Marquard-Algorithmus optimiert werden. Die zu minimierende Kostenfunktion

$$E(P) = \sum_{\forall i} d\left(\mathbf{x}_{ij}, P_j \mathbf{X}_i\right)^2 . \tag{5.47}$$

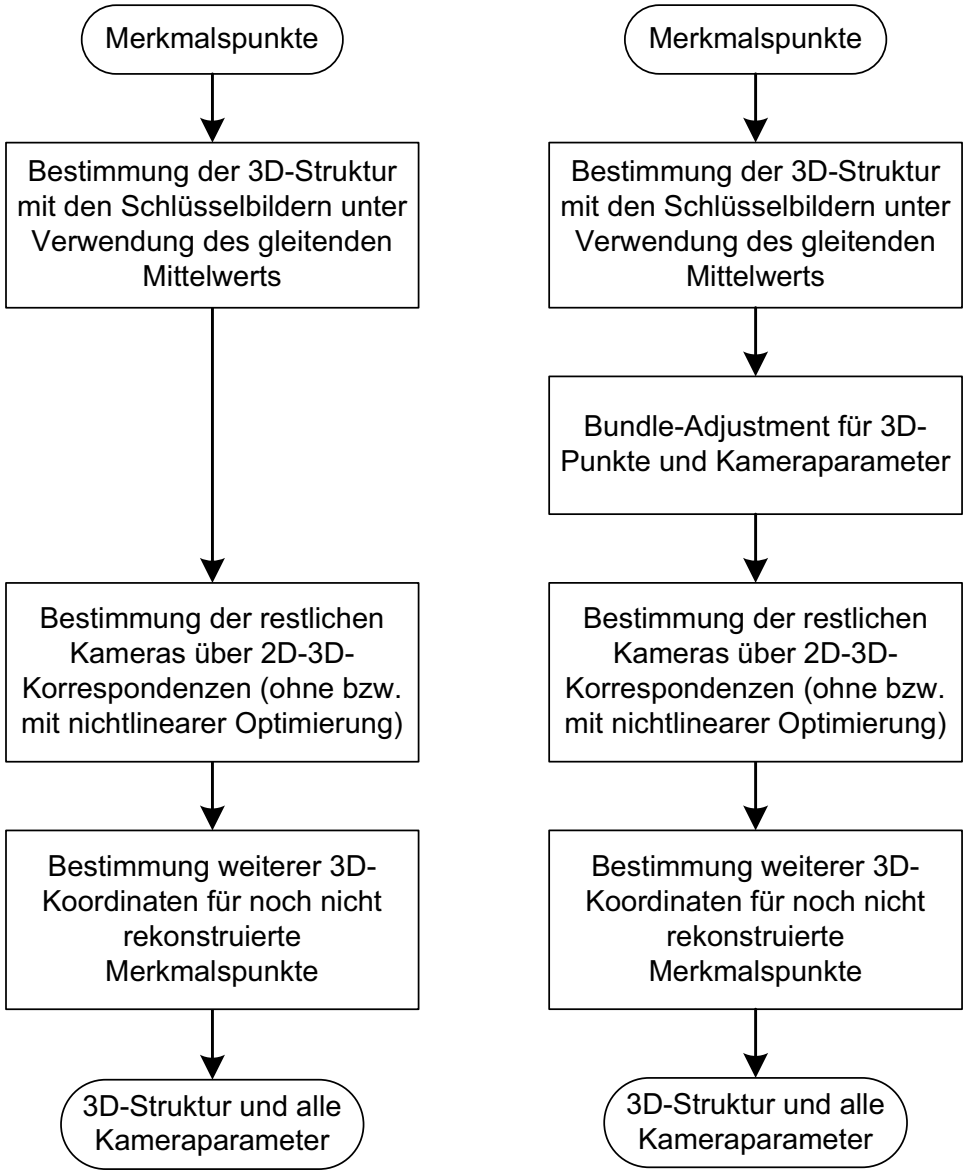


Abbildung 5.11.: Bestimmung der Kameraparameter und 3D-Struktur: (a) ohne und (b) mit Bundle Adjustment

ist dabei nur von den zwölf Parametern der aktuellen Projektionsmatrix P_j abhängig. Der Ablauf der Rechenschritte mit nichtlinearer Erweiterung ist ebenfalls in Abbildung 5.11(a) dargestellt.

Nichtlineare Rekonstruktion mithilfe von Bundle Adjustment für die Schlüsselbilder

Es wird nun Bundle Adjustment zur nichtlinearen Optimierung sowohl der 3D-Struktur als auch der Projektionsmatrizen verwendet. Nach der Rekonstruktion mithilfe der Schlüs-

selbilder soll die Kostenfunktion in Gl. (5.46) ohne die Zwischenansichten minimiert werden, d.h. die Anzahl der zu optimierenden Parameter ist durch die geringere Anzahl an Ansichten reduziert. Erst im Anschluss an das Bundle Adjustment werden die Projektionsmatrizen der Zwischenansichten über 2D-3D-Punktkorrespondenzen ermittelt. In Abbildung 5.11(b) ist der Ablauf der Rechenschritte um das Modul „Bundle Adjustment“ erweitert.

Nichtlineare Rekonstruktion mithilfe globaler Optimierung

Die letzte Implementierungsvariante, die verglichen werden soll, ist die der globalen nichtlinearen Optimierung für die gesamte Sequenz. D.h. Bundle Adjustment wird zuletzt für sämtliche Projektionsmatrizen und 3D-Punkte durchgeführt, indem Gl. (5.46) inklusive Zwischenansichten minimiert wird.

5.5.5. Rekonstruktionsergebnisse

Die vorgestellten Implementierungsvarianten sollen nun untereinander verglichen werden. Als Videodatensatz wurde die Sequenz *Statue* mit 210 Einzelbildern[3] und einer Auflösung von 720×576 Pixeln verwendet. Drei Beispielbilder der Videosequenz sind in Abbildung A.13 im Anhang A.3 dargestellt. Insgesamt wurden 400 Merkmalspunkte über die Sequenz verfolgt.

Evaluierung 1: Zuerst wird die lineare Rekonstruktionsmethode ohne und mit Bundle Adjustment nach Gl 5.46 für die Schlüsselbilder untersucht. Der Verlauf der mittleren Reprojektionsfehler für beide Implementierungsvarianten ist in Abbildung 5.12 dargestellt. Durch das Bundle Adjustment konnten große Ausschläge des Fehlers (z. B. bei den Bildern 23 und 31) stark reduziert werden. Allerdings hat sich der Fehler vereinzelt auch leicht erhöht, da durch das Bundle Adjustment der quadratische Fehler über alle Ansichten minimiert wird, d.h. große Fehler haben einen stärkeren Einfluss auf das Nachregeln der Parameter für die Projektionsmatrizen und die 3D-Punkte, wodurch sich zuvor kleine Fehler etwas vergrößern können. Der mittlere Reprojektionsfehler über alle 210 Ansichten der Sequenz hat sich von $0,4724$ Pixel (linear) auf $0,4436$ Pixel (linear + BA) durch das Bundle Adjustment reduziert.

[3] Es wurde hier nur jedes dritte Bild der Sequenz mit insgesamt 630 Bildern ausgewählt.

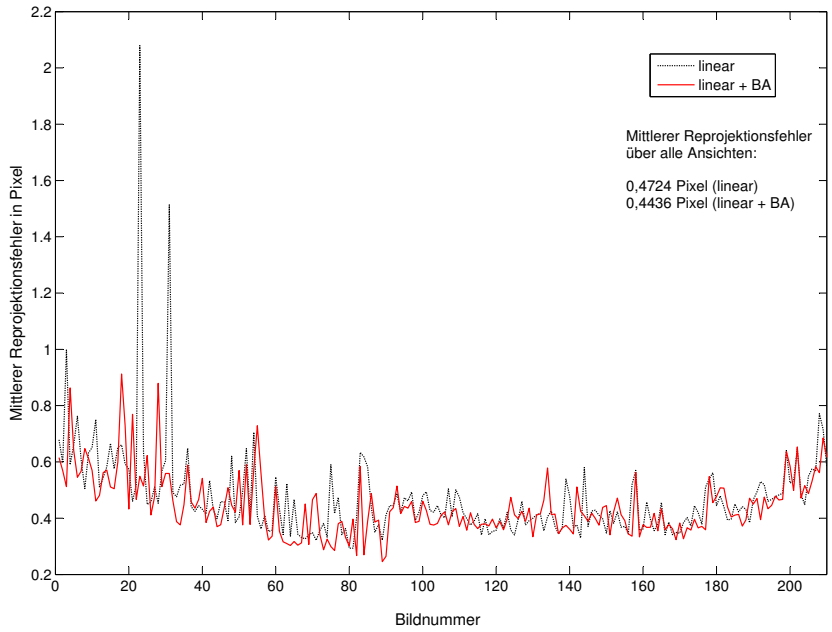


Abbildung 5.12.: Vergleich des mittlerer Reprojektionsfehlers für lineare Berechnung der Projektionsmatrizen ohne und mit Bundle Adjustment für die Schlüsselbilder

Evaluierung 2: In einer zweiten Untersuchung wurden die Parameter der Projektionsmatrizen, die über 2D-3D-Korrespondenzen ermittelt werden, mit Hilfe des LM-Algorithmus nichtlinear optimiert. Entsprechend der Evaluierung 1 wird diese Implementierungsvariante der Variante mit Bundle Adjustment für die Schlüsselbilder gegenübergestellt. Abbildung 5.13 stellt die Verläufe der mittleren Reprojektionsfehler für jedes Bild der Sequenz dar. Die Verringerung des Fehlers durch das Bundle Adjustment fällt hier geringer aus. Der mittlere Reprojektionsfehler über alle 210 Ansichten hat sich von $0,5013$ Pixel (nichtlinear P) auf $0,4996$ Pixel (nichtlinear P + BA) reduziert. Zwar liegen diese Werte über den Werten der linearen Rekonstruktionsmethode aus Evaluierung 1, ein direkter Vergleich soll aber nicht vorgenommen werden, da durch die Verwendung von RANSAC - vor allem in der Initialisierungsphase - eine neue Berechnung der Parameter zu leicht unterschiedlichen Ergebnissen führen kann. Ein Vergleich der Verläufe für die lineare und die nichtlineare Rekonstruktionsmethode kann hingegen vorgenommen werden. Hier wird schnell deutlich, dass die nichtlineare Berechnung der Projektionsmatrizen einen signifikant glatteren Verlauf zu Folge hat als die lineare Rekonstruktionsmethode.

Evaluierung 3: Abschließend wurde auf die Ergebnisse der nichtlinearen Rekonstruktionsmethode mit Bundle Adjustment für die Schlüsselbilder eine globale Optimierung mittels Bundle Adjustment für die Parameter aller Projektionsmatrizen und 3D-Punkte

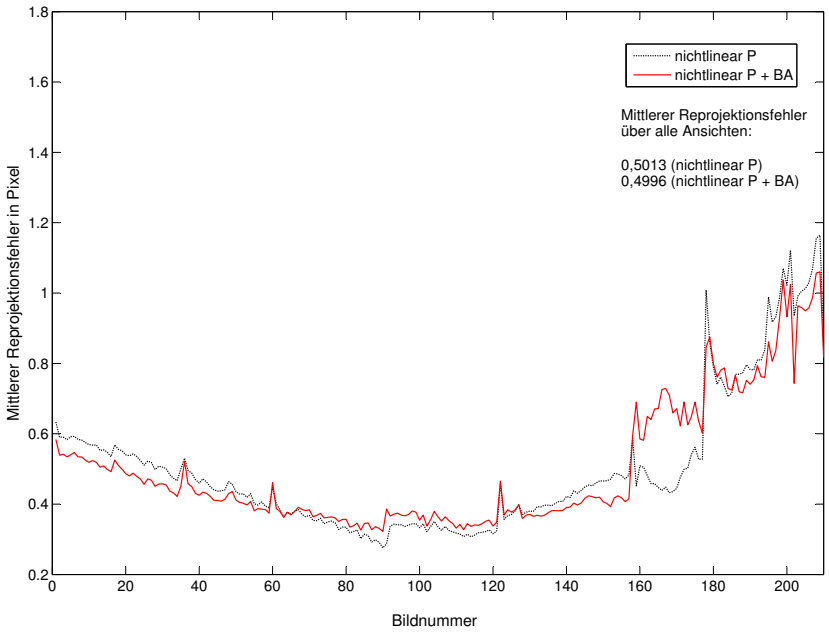


Abbildung 5.13.: Vergleich des mittlerer Reprojektionsfehlers für nichtlineare Berechnung der Projektionsmatrizen ohne und mit Bundle Adjustment für die Schlüsselbilder

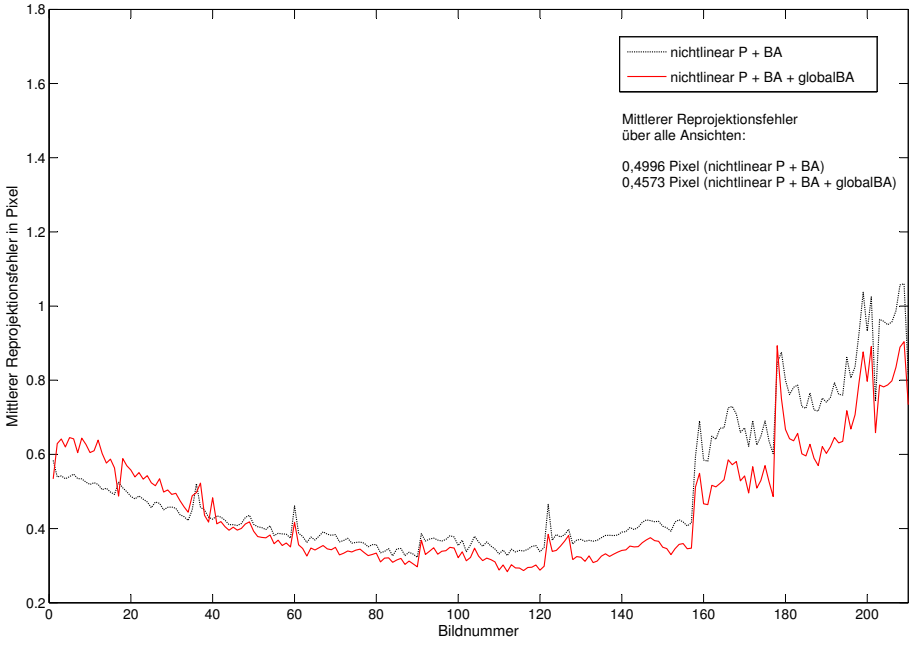


Abbildung 5.14.: Vergleich des mittlerer Reprojektionsfehlers für nichtlineare Berechnung der Projektionsmatrizen mit Bundle Adjustment für die Schlüsselbilder und globaler Optimierung

durchgeführt. Abbildung 5.14 stellt die Verläufe der mittleren Reprojektionsfehler grafisch dar. Durch die globale Optimierung konnte der mittlere Fehler über alle 210 Ansichten von 0,4996 Pixel (nichtlinear P + BA) auf 0,4573 Pixel (nichtlinear P + BA + globalBA) deutlich reduziert werden. Die Verringerung tritt vor allem zwischen Bild 158 und 210 auf, da hier der Reprojektionsfehler einen größeren Anstieg zu verzeichnen hat.

5.6. Weitere Rekonstruktionsmethoden

Aus den vorherigen Betrachtungen wird schnell klar, dass sich weitere Kombinationen von linearen und nichtlinearen Optimierungsansätzen bzw. robusten und nichtlinearen Ansätzen bilden lassen. Einige weitere Varianten wurden implementiert, sollen aber aufgrund von geringfügigen Änderungen der Ergebnisse in dieser Arbeit nicht weiter verfolgt werden. Für interessierte Leser wird auf die Studienarbeit von Zhao [204] verwiesen. Einige abweichende Rekonstruktionsmethoden sollen aber an dieser Stelle noch erwähnt und kurz beschrieben werden.

In [84] wird ein Verfahren vorgestellt, welches die Sequenz in zeitliche Abschnitte unterteilt (z. B. durch Detektion von Schlüsselbildern), diese aber vorerst unabhängig voneinander behandelt. D.h. zuerst wird eine 3D-Rekonstruktion der einzelnen zeitlichen Abschnitte (engl. *temporal window*, TW) TW_1 bis TW_N mit N als Anzahl der zeitlichen Abschnitte durchgeführt. Anschließend werden benachbarte Abschnitte über eine 3D-Homographie zusammengeführt, d.h. die Kameraprojektionsmatrizen und die berechneten 3D-Punkte werden in ein gemeinsames Weltkoordinatensystem transformiert. Daraus entstehen wiederum zeitliche Abschnitte TW_{1-2}, TW_{3-4} usw., die ebenfalls im nächsten Schritt wieder zusammengeführt werden. Im letzten Schritt gelangt man zu der finalen Rekonstruktion TW_{1-N}, welche mithilfe eines globalen Optimierungsverfahren - z. B. Bundle Adjustment - optimiert werden kann. Eine detaillierte Vorgehensweise zum Zusammenführen zweier Teilrekonstruktionen und damit zur Bestimmung einer 3D-Homographie ist in [26] beschrieben.

Diese Art der Rekonstruktion wird auch als *hierarchische* 3D-Rekonstruktion bezeichnet. Der Vorteil hierarchischer Rekonstruktionen liegt in der unabhängigen Rekonstruktion von Teilabschnitten. Rekonstruktionsfehler, die bei sequentiellen Ansätzen in der initialen Rekonstruktion auftreten, pflanzen sich nicht im Verlauf der Rekonstruktion fort. Etwas schwieriger gestaltet sich bei hierarchischen Ansätzen allerdings die Bestimmung von 3D-Homographien, da die Rekonstruktionsergebnisse zweier benachbarter Teilrekonstruktionen im dreidimensionalen Raum zusammengeführt werden müssen.

Zu den sogenannten *Batch*-Methoden zählt die *Faktorisierung*, die in [179] vorgestellt und in [62] für verschiedene Kameramodelle erweitert wurde. Bei dieser Methode werden

sämtliche Merkmalspunkttrajektorien in einer Matrix zusammengefasst und über Singulärwertzerlegung in zwei Untermatrizen, die zum einen die 3D-Koordinaten der Punkte und zum anderen die Kameraparameter enthält, faktorisiert.

Ein weiteres Verfahren, welches vorerst den sequentiellen Ansatz nach Pollefeys [140] für eine initiale Rekonstruktion der gesamten Sequenz verwendet und anschließend einen hierarchischen Ansatz, um die Rekonstruktionsergebnisse zu optimieren, wird in [76] und [77] vorgestellt. In einem ersten Schritt werden dabei die 3D-Struktur und die Kameramatrizen mithilfe des beschriebenen sequentiellen Ansatzes bestimmt. Da es sich hierbei um eine initiale Rekonstruktion handelt, sind nichtlineare Optimierungsverfahren bezüglich der Rechenkomplexität nicht notwendig. In einem nächsten Schritt werden neue Schlüsselbildpaare ermittelt, die sowohl einen möglichst großen Abstand der Kamerazentren als auch eine große Anzahl an 2D-Punktkorrespondenzen haben. Für diesen Zweck wird ein Gewichtungsmaß definiert

$$p = d + \frac{\alpha}{1 + e^{\beta(n-\gamma)}}, \tag{5.48}$$

welches die Nutzbarkeit eines Bildpaares für die 3D-Rekonstruktion bestimmt. Dabei ist d der Abstand der Kamerazentren und n die Anzahl an Punktkorrespondenzen zwischen den beiden Ansichten. α, β und γ sind Parameter einer Sigmoid-Funktion, die den zweiten Term in Gl. (5.48) nichtlinear gewichten sollen.

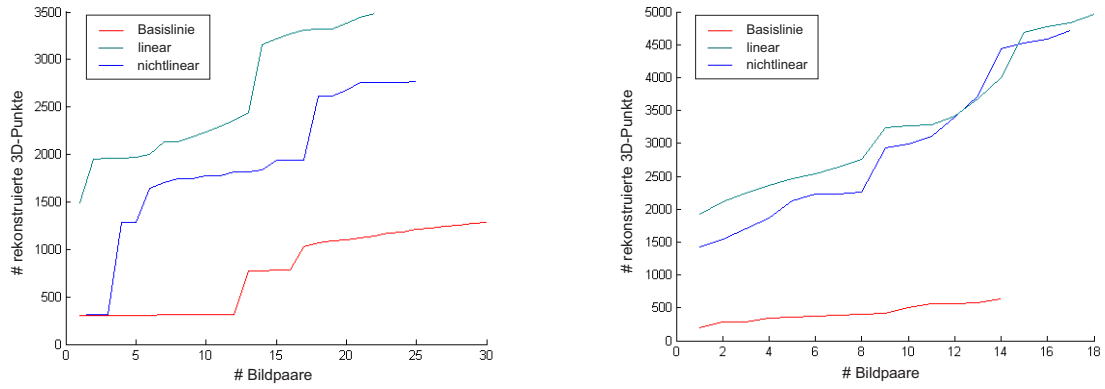
Mit den neu sortierten Bildpaaren (beginnend mit dem Bildpaar, für das p den größten Wert liefert), wird nun der hierarchische Ansatz verfolgt, indem 3D-Rekonstruktionen nacheinander für die Bildpaare durchgeführt werden. Diese Teilrekonstruktionen sind nur dann unabhängig voneinander, solange noch keine der beiden Ansichten eines Bildpaares für eine Rekonstruktion verwendet wurde.

In Tabelle 5.4 sind die Auswirkungen von drei verschiedenen Gewichtungsmaßen auf die Rekonstruktionsergebnisse dargestellt. Zum einen wird nur der Abstand der Kamerazentren („Basislinie“) berücksichtigt (erster Term in Gl. (5.48)) und zum anderen zusätzlich die Anzahl der Punktkorrespondenzen, wobei hier zwischen „linear“ und „nichtlinear“ unterschieden wird. Für den ersten Fall wurden die Länge der Basislinie und die Anzahl der Punktkorrespondenzen linear gewichtet. Der nichtlineare Fall entspricht dem Gewichtungsmaß in Gl. (5.48).

Zwar liefert „Basislinie“ den geringsten mittleren Reprojektionsfehler, die Anzahl an insgesamt rekonstruierten 3D-Punkten und damit die Glaubwürdigkeit der Rekonstruktionsergebnisse ist aber mit Abstand am geringsten. Das nichtlineare Gewichtungsmaß zeigt dabei einen sehr guten Kompromiss hinsichtlich Reprojektionsfehlern und Anzahl an 3D-Punkten. An dieser Stelle sei noch vermerkt, dass die Anzahl der möglichen Bildpaarungen 28680 bzw. 19900 für die synthetische Sequenz *TUBroom* (Auflösung: 720×576, 240 Einzelbilder) bzw. für die TV-Sequenz *Palast* beträgt. Die Anzahl der rekonstruierten

Tabelle 5.4.: Auswirkungen verschiedener Gewichtungsmaße auf die Rekonstruktionsergebnisse (Quelle: [77])

	Anzahl an Bildpaaren	Anzahl rekonstruierter 3D-Punkte	Mittlerer Reprojektionsfehler
TUBroom			
Basislinie	14	630	0,19
Linear	18	4961	0,58
Nichtlinear	17	4716	0,27
Palast			
Basislinie	30	1284	0,73
Linear	22	3481	1,31
Nichtlinear	25	2771	0,9



(a) (b)

Abbildung 5.15.: Vergleich verschiedener Gewichtungsmaße zur Bestimmung neuer Bildpaare: (a) TV-Sequenz *Palast*, (b) synthetische Sequenz *TUBroom* (Quelle: [77]).

3D-Punkte nach der Initialisierungsphase betrug 6095 (*TUBroom*) bzw. 3546 (*Palast*). Daraus folgt, dass sich mit sehr wenigen gut gewählten Bildpaaren die 3D-Struktur bestimmen lassen kann.

Die Abbildungen 5.15(a) und 5.15(b) zeigen den Verlauf der rekonstruierten 3D-Punkte im Verhältnis zur Anzahl der verwendeten Bildpaare.

5.7. Ergebnisse

In den Abbildungen 5.16 und 5.17 sind die Rekonstruktionsergebnisse für die Videosequenzen *Kirche* (Auflösung: 576×720, 555 Einzelbilder) und *Dom* (Auflösung: 720×576,

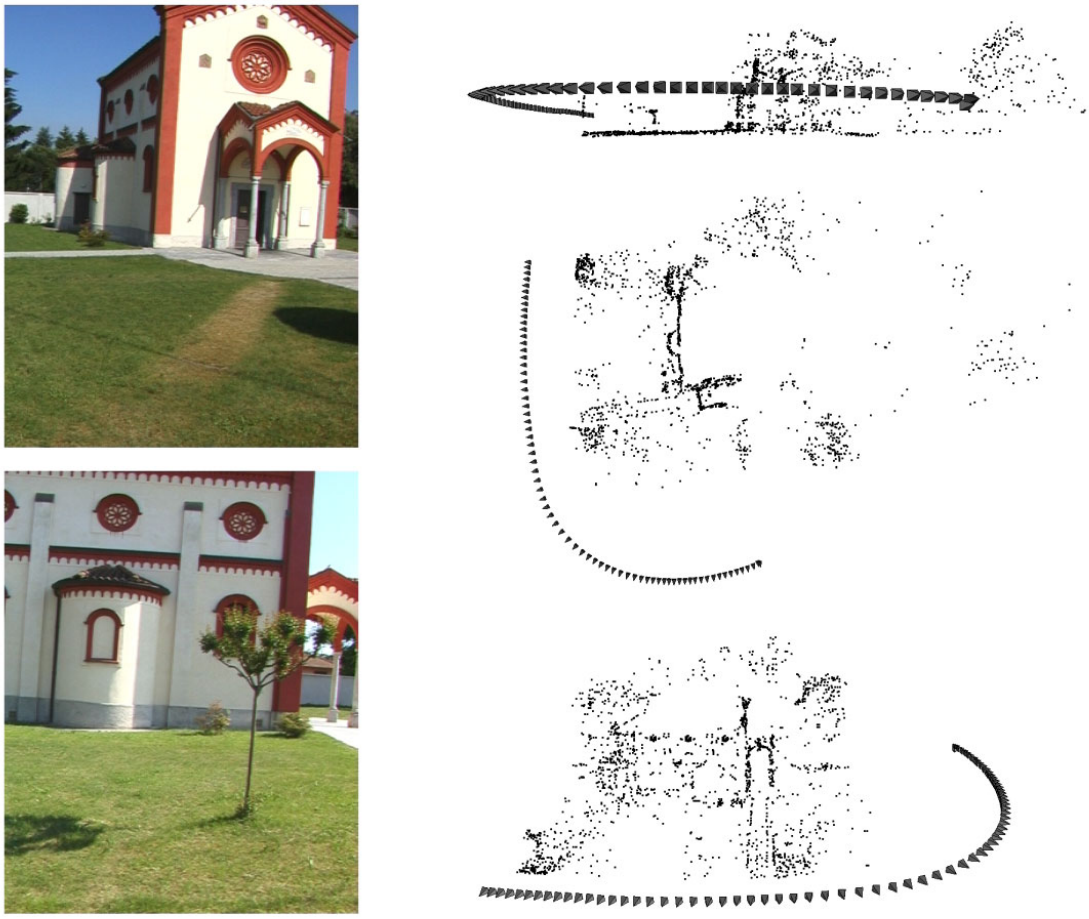


Abbildung 5.16.: Rekonstruktionsergebnis der Sequenz *Kirche*

1400 Einzelbilder) dargestellt. Auf der linken Seite der Abbildungen werden jeweils zwei Beispielbilder der Videosequenzen gezeigt und auf der rechten Seite entsprechend die 3D-Rekonstruktion der Szenen mit den Kamerapositionen und -orientierungen aus unterschiedlichen Perspektiven.

5.8. Kapitelzusammenfassung

In diesem Kapitel wurden verschiedene Aspekte der 3D-Rekonstruktion aus Videosequenzen mit bewegter Kameraführung behandelt. Zum einen ist die Funktionsweise von Verfahren bei der Verfolgung und Überwachung von Merkmalspunkten umfangreich erläutert worden und zum anderen wurde die Bestimmung der Kameraparameter und 3D-Struktur anhand eines sequentiellen Ansatzes beschrieben.

Bei der Verfolgung und Überwachung von Merkmalspunkten wurden verschiedene Kostenfunktionen (SSD, NCC) und Bewegungsmodelle (translatorisch, affin) untersucht

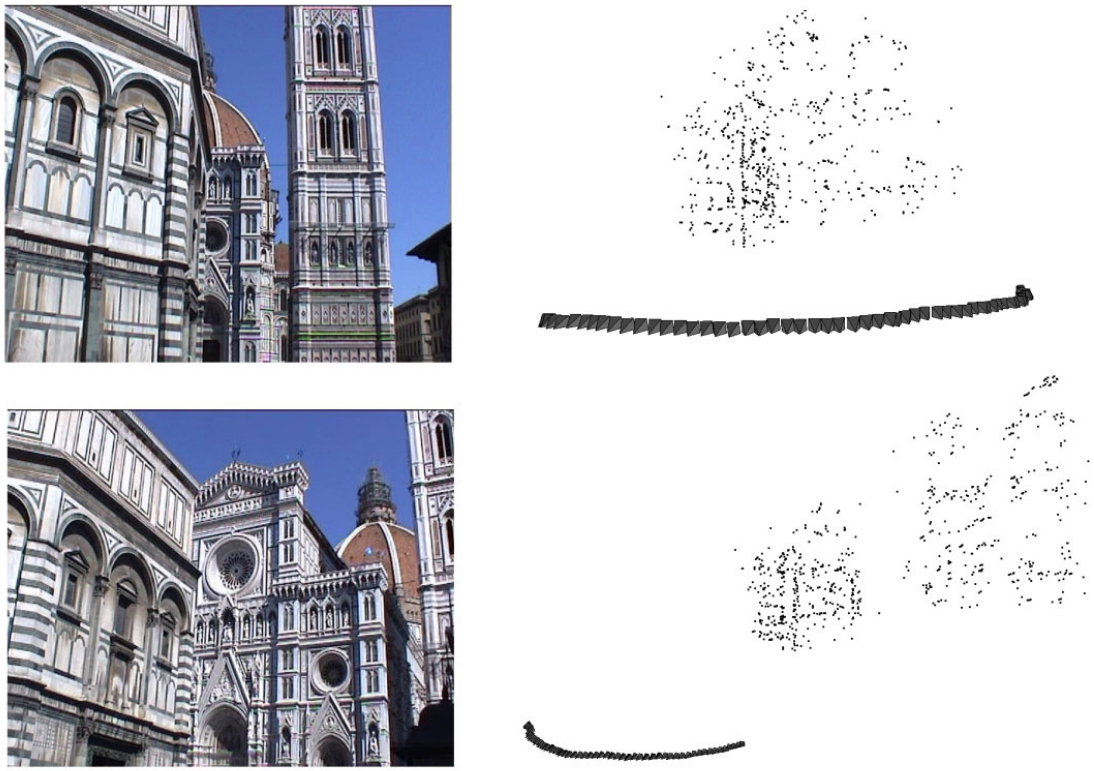


Abbildung 5.17.: Rekonstruktionsergebnis der Sequenz *Dom*

und deren Vor- und Nachteile erörtert. Die Untersuchungen ergaben, dass für den Überwachungsprozess der verfolgten Merkmalspunkte die NCC-basierte Kostenfunktion in Kombination mit dem affinen Bewegungsmodell besonders geeignet ist. Da die Berechnung der Deformationsmatrix des affinen Bewegungsmodells sehr rechenintensiv ist, wurden zusätzliche Untersuchungen angestellt, um zu prüfen, in wie weit die Berechnung der Deformationsmatrix überhaupt sinnvoll ist. Durch die Festlegung zweier Schwellwerte, zwischen denen das affine Bewegungsmodell lediglich berechnet wurde, konnte die Rechenkomplexität erheblich reduziert werden, ohne einen Qualitätsverlust beim Überwachungsprozess hinnehmen zu müssen.

Für die Bestimmung der Kameraprojektionsmatrizen und der 3D-Struktur wurde ein sequentieller Ansatz vorgestellt, welcher durch Kombination von linearen und nichtlinearen bzw. globalen Optimierungsverfahren näher untersucht wurde. Ein Problem der Rekonstruktion, der geringe Abstand der Kamerazentren benachbarter Bilder einer Videosequenz, wurde über die Berechnung zweier Bewegungsmodelle (H und F) gelöst, indem Schlüsselbilder über das geometrisch robuste Informationskriterium (GRIC) ausgewählt wurden. Die untersuchten Implementierungsvarianten der sequentiellen Rekonstruktion zeigten, dass sich die Reprojektionsfehler durch nichtlineare und globale Optimierungsverfahren zum Teil erheblich reduzieren ließen. Auf die zusätzliche Komplexität der Re-

chenoperationen speziell für das verwendete Bundle Adjustment wurde nur indirekt eingegangen, da dies bereits hinlänglich bekannt ist.

Ferner wurde noch ein hierarchischer Rekonstruktionsansatz vorgestellt, der auf Basis einer initialen Rekonstruktion eine erneute Sortierung der Bilder vornimmt, um die bestmöglichen Schlüsselbildpaarungen zu ermitteln und damit die Rekonstruktionsergebnisse zu verbessern.

6. Generierung stereoskopischer Ansichten

Die zuvor beschriebenen Kapitel behandelten vorrangig Verfahren zur Analyse von Bildern und Videosequenzen, welche die 3D-Informationen, die in dem zugrunde liegenden Bildmaterial vorhanden sind, extrahieren. In diesem Kapitel soll nun auf Basis dieser 3D-Informationen das Augenmerk auf die Synthese neuer virtueller Ansichten gelegt werden. Diesbezüglich wird ein neues Verfahren zur Stereobildsynthese vorgestellt, das ausschließlich mit 2D-Transformationen arbeitet. In dem folgenden Abschnitt wird zunächst ein Verfahren zur Erzeugung von stereoskopischen Sequenzen beschrieben, welches zusätzlich zu den Bilddaten die dichte Tiefeninformation für jedes Pixel der einzelnen Bilder nutzt. Dieses Verfahren, das zur Standardisierung in MPEG-4 Part 16: Animation Framework eXtension (AFX) aufgenommen wurde [80], gehört zu dem aktuellen Stand der Forschung und ist unter dem Namen *Depth Image-based Rendering* (DIBR) bekannt.

Eine aussagefähige Evaluierung zum Vergleich der beiden vorgestellten Ansätze ist in Kapitel 7 zu finden.

6.1. Stereo aus Video und Tiefe

DIBR gehört derzeit zu den am meisten verbreiteten Ansätzen zur Generierung stereoskopischer Ansichten [8, 22, 27, 42–45, 63, 87, 88, 95, 173, 175, 198, 199], da es eine Reihe von Vorteilen bei der Kodierung, Übertragung und Darstellung in 3D-Anwendungen liefert. Voraussetzung für die Erzeugung stereoskopischer bzw. neuer virtueller Ansichten ist, dass für jedes Bild einer Bildfolge die entsprechende Tiefeninformation jedes einzelnen Pixels bekannt sein muss. Diese Tiefeninformation wird in Form einer Tiefenkarte (engl. *depth map*) durch Grauwerte dargestellt. In der Regel werden entferntere Objekte in einem Bild durch dunklere Grautöne repräsentiert. Abbildung 6.1 zeigt ein Beispiel eines Videobildes mit zusätzlicher Tiefenkarte.

6.1.1. Problemstellung und Überblick über die Methode

Das Ziel von DIBR ist es, aus einer Originalansicht mithilfe der Tiefeninformation für jedes Pixel eine neue virtuelle Ansicht zu generieren, die die aufgenommene 3D-Szene von einer zweiten virtuellen Perspektive betrachtet und damit die Aufnahme einer zweiten Kamera simuliert. Da Objekte entsprechend ihrer Lage im dreidimensionalen Raum

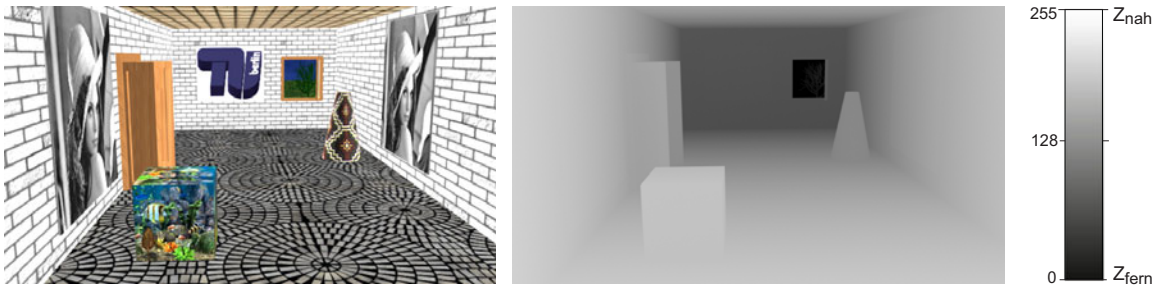


Abbildung 6.1.: Darstellung von Video und Tiefe: Originalbild und Tiefenkarte der synthetischen Sequenz *TUBroom*

unterschiedlich versetzt in zwei Kameras abgebildet werden, kann die Tiefeninformation genutzt werden, die Pixel entsprechend ihrer Tiefe verschieden stark in einer virtuellen Ansicht zu verschieben. Der Versatz von Szenenobjekten innerhalb von zwei unterschiedlichen perspektivischen Ansichten wird als *Parallaxe* bezeichnet. Bei einer Aufnahmeeinrichtung, bei der die Kameras z. B. parallel angeordnet sind, werden Pixel, die zu einem nahe liegenden Objekt gehören, stärker versetzt als Pixel die zu dem weit entfernten Hintergrund gehören, d.h. sie besitzen eine stärkere Parallaxe.

Allerdings werden dadurch Bereiche, die in der Originalansicht durch Vordergrundobjekte noch verdeckt waren, in der virtuellen Ansicht aufgedeckt. Die Problematik der Auf- und Verdeckung von sichtbaren und nicht sichtbaren Bereichen spielt vor allem in der Computer Graphik (CG) [121,123] eine große Rolle. Da keinerlei Informationen dieser aufgedeckten Bereiche sowohl in der Originalansicht als auch in der Tiefenkarte vorhanden sind, treten vorerst sogenannte *Löcher* (Pixel ohne Textur- bzw. Farbinformation) auf. Um die Löcher mit Farbinformationen zu füllen und somit zu verschleiern, werden Verfahren wie die Extrapolation der Hintergrundinformation, die Interpolation von Vorder- und Hintergrund oder die Spiegelung des Hintergrunds angewendet. Eine weitere Technik, um Aufdeckungen bzw. Löcher zu vermeiden, wurde von Fehn [44] vorgestellt, indem eine Vorverarbeitung der Tiefenkarte durch Gaußfilterung vorgenommen wird.

Ein zusätzliches, wenn auch geringeres, Problem bei dem Versatz von Pixeln ist die Tatsache, dass diese nicht unmittelbar auf dem quantisierten Pixelraster der virtuellen Kamera liegen. Da der Pixelversatz allerdings lediglich in horizontaler Richtung vorgenommen wird, können eindimensionale Interpolationsverfahren eingesetzt werden. Interpolationen besitzen allerdings eine Tiefpasscharakteristik, d.h. Regionen mit hohem Texturgehalt und vor allem Kanten werden unschärfer dargestellt.

Ein weitaus größeres Problem stellt jedoch die Gewinnung der Tiefeninformation dar. DIBR ist lediglich ein Bildsyntheseverfahren, das auf die Textur- und Tiefeninformation zurückgreift. Bei der Aufnahme einer Szene mit einer herkömmlichen Kamera wird nur die Textur auf der Bildebene abgebildet. Um die Tiefe für jedes Pixel des Bildes zu erhalten

müssen aktive (z. B. Laser-Range-Scanner) oder passive Verfahren zur Erzeugung dichter Tiefenkarten herangezogen werden.

6.1.2. Erzeugung dichter Tiefenkarten

Die Erzeugung dichter Tiefenkarten ist ein großes Forschungsfeld in der Computer Vision. Generell wird dabei zwischen aktiven und passiven Verfahren zur Tiefenbestimmung unterschieden. Aktive Verfahren sind für die Zielsetzung dieser Arbeit ungeeignet, da für diese eine zusätzliche Aufnahmeeinrichtung - wie z. B. ein Laser-Range-Scanner [13, 56, 58, 163] oder eine sogenannte Zcam [74] - zur Messung der Tiefe erforderlich ist. Sind nur die Bildinformationen vorhanden, müssen die Tiefenkarten durch passive Verfahren ermittelt werden. Die passiven Verfahren können dabei noch einmal in Verfahren unterteilt werden, die monokulare Techniken (also lediglich die Tiefeniformationen aus einem Bild entnehmen) verwenden, und Methoden, die auf mehreren Ansichten einer Szene basieren (binokulare Techniken), wie sie in den Kapiteln 4 und 5 vorgestellt wurden.

Zu den monokularen Tiefenschätzverfahren zählen

- Tiefe aus Fokus/Defokus
- Tiefe aus Textur und
- Tiefe aus Schattierung.

Verfahren wie Tiefe aus Fokus (engl. *Shape-from-Focus*, SfF) und Tiefe aus Defokus (engl. *Shape-from-Defocus*, SfD) versuchen durch geänderte Kameraoptik 3D-Objektoberflächen einer Szene anhand der Bildschärfe zu rekonstruieren. Beide Verfahren benötigen zwar mehrere Bilder als Basis für die Tiefenschätzung, sie werden aber dennoch zu den monokularen Tiefenschätzverfahren gezählt, da sich die externen Parameter der Kamera nicht ändern. Als Kriterium für die Beurteilung der Bildschärfe dient bei den meisten Anwendungen das Verhältnis der hochfrequenten zu den niederfrequenten Anteilen im Bild, da z. B. die Defokussierung eine Tiefpasscharakteristik aufweist. Bei SfF werden gezielt die Kameraparameter wie fokale Länge und Öffnung der Blende variiert und nach Objekten gesucht, die sich im Fokus befinden bzw. eine maximale Schärfe besitzen. Umgekehrt wird bei SfD die Stärke der Unschärfe als Maß für relative Tiefenunterschiede untersucht. Zwar wird bei SfD-Verfahren ebenfalls mit veränderter Kameraoptik gearbeitet, diese muss allerdings nicht bekannt sein. Allgemein werden bei SfD-Verfahren weniger Bilder mit variierten Kameraparametern benötigt als bei SfF. Allerdings ist SfD in der Regel ungenauer als SfF und wird aufgrund der schnelleren Verarbeitungszeiten als Vorstufe von SfF Verfahren eingesetzt [172]. Ein guter Überblick über SfF und SfD ist in den Dissertationen von Wei [194] und Liu [111] sowie in [41] zu finden.

Hinweise auf Objektformen, Objektgrößen und Entfernungen sind u.a. auch direkt in den Texturinformationen eines Bildes zu finden. Verfahren, die die Tiefe aus der Textur (engl. *Shape-from-Texture*, SfT) berechnen, sind besonders wirkungsvoll, wenn die genaue Form der elementaren Texturelemente (*Texel*) im Vorfeld bekannt ist [86]. Unter Berücksichtigung der perspektivischen Projektion kann aus der Deformierung und der Größenänderung der Texel die Orientierung der Objektoberfläche bestimmt werden. Für viele Aufnahmen ist die Form der Texel in der Regel nicht bekannt. Durch regelmäßig auftretende Muster können diese aber ermittelt werden und somit Rückschlüsse auf die Oberflächenorientierung liefern [15]. Diese Ansätze werden auch als *strukturelle Texturanalyse* bezeichnet. Allerdings treten periodische Muster in der Natur eher selten auf, so dass hierfür Texturmerkmale wie Gradient, Entropie, Korrelation etc. extrahiert werden müssen, weshalb diese Art der Tiefengewinnung aus Textur auch als *statistische Texturanalyse* bezeichnet wird. Da die statistische Texturanalyse sehr komplex und gleichzeitig unzuverlässig ist, findet sie praktisch nur in der Mustererkennung oder Bildsegmentierung Anwendung (z. B. in [69, 73]). Geometrische Szeneneigenschaften wie Fluchtpunkte und Fluchtlinien lassen sich ebenfalls mithilfe der Texturanalyse durch Extrahieren von Kanten ermitteln [70, 105, 137].

Bei den schattierungsbasierten Verfahren (engl. *Shape from Shading*, SfSh) werden die Unterschiede der Grauwerte im Bild direkt in Beziehung zu Oberflächenorientierungen in der Szene und der Beleuchtung gesetzt. Ein einfaches Modell zur Beschreibung der Beziehungen zwischen Lichtquelle, Oberfläche und Bildpunkt ist das Lambertsche Modell, bei dem der Grauwert eines Pixels von der Richtung der Lichtquelle und der Richtung der Oberflächennormalen abhängt [146]. Sind die Reflexionseigenschaften einzelner Objektoberflächen bekannt (z. B. linear reflektierende Oberflächen) und das Rückstrahlungsvermögen der Objekte konstant, so lassen sich die relativen Tiefenwerte durch globale Minimierungs- oder Ausbreitungsansätze, wie die Bestimmung der Oberfächeneigenschaften durch Minimierung einer Energiefunktion oder durch Propagieren der Oberflächeninformationen ausgehend von bekannten Oberflächenpunkten über das gesamte Bild, bzw. durch lokale Ansätze, die nur die Strahlungsintensitäten in nächster Nähe eines Oberflächenpunktes betrachten, ermitteln [97]. Ein guter Überblick und Vergleich der gängigsten SfSh-Verfahren ist in Zhang et al. [200] und Durou et. al [33] zu finden.

Zu den Tiefenschätzverfahren, die zwei oder mehrere Ansichten aus verschiedenen Perspektiven verwenden, zählen

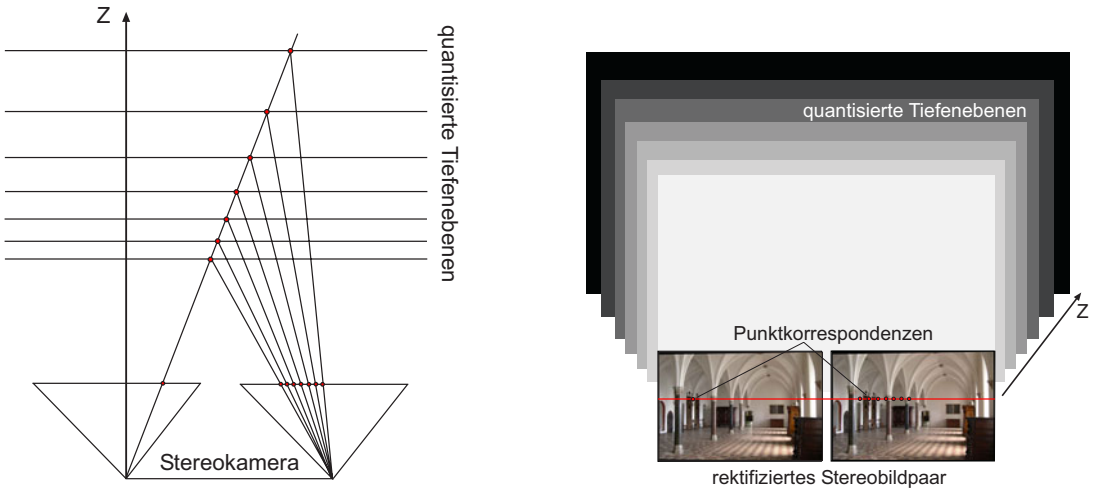
- Tiefe aus Stereo
- Tiefe aus Bewegung und
- Tiefe bzw. Form aus Silhouetten.

Die Verfahren Tiefe aus Stereo (engl. *Shape from Stereo* bzw. *Structure from Stereo*, SfS) und Tiefe aus Bewegung (engl. *Structure from Motion*, SfM) werden in der Literatur häufig gleich gesetzt, da sie vom Prinzip der Tiefenschätzung auf identischen Algorithmen basieren. Um die Unterschiede klar hervorzuheben, soll an dieser Stelle darauf hingewiesen werden, dass die Szene bei SfS-Verfahren mit zwei oder mehreren Kameras synchron aufgenommen wurde. Bei SfM wurde lediglich eine Kamera verwendet, die sich während der Aufnahme zumindest translatorisch bewegt hat. Ein weiterer wesentlicher Unterschied beider Verfahren ist, dass SfS gegenüber SfM je nach verwendetem Algorithmus die Tiefe in Echtzeit berechnen kann. Bei SfM ist es schon rein physikalisch nicht möglich, dichte Tiefenkarten innerhalb der Bildrate zu ermitteln, da erst nach einer hinreichend großen Kamerabewegung, also nach der Aufnahme einer entsprechenden Bildfolge, eine Tiefenschätzung durchgeführt werden kann (siehe Abschnitte 5.1 und 5.3).

SfS-Verfahren basieren auf Disparitätsschätzungen für jedes Pixel innerhalb eines Bildes. Da durch die Verwendung mehrerer Kameras die 3D-Szene von verschiedenen Perspektiven aufgenommen wird, werden sämtliche 3D-Raumpunkte entsprechend ihrer Tiefe an unterschiedlichen Positionen in den Bildebenen der Kameras abgebildet. Die Disparität, also die Verschiebung der Punkte innerhalb der Bildebene, gibt Aufschluss über die relativen Tiefen. Durch Triangulationsverfahren (siehe Abschnitt 4.6) lassen sich die 3D-Koordinaten der Punkte ermitteln.

SfS zählt zu den am weitest verbreiteten passiven Tiefenschätzverfahren und wird aufgrund der Genauigkeit und der geringen Rechenzeiten häufig in Anwendungsgebieten wie z. B. der Roboter Vision [19, 150, 191, 203] oder der immersiven Telekonferenzsysteme [5, 36, 131, 189] eingesetzt. SfS-Verfahren besitzen den Vorteil, dass sie auf eine Reihe von Einschränkungen bei der Disparitätsschätzung zurückgreifen können. In der Regel unterliegen Stereokameras einem festen Setup, d.h. der Abstand und die Orientierung der Kameras zueinander bleibt bei der Aufnahme unverändert. Somit kann bereits vor der Aufnahme z. B. die Fundamentalmatrix berechnet und eine Kalibrierung vorgenommen werden [203], so dass durch einmalige Rektifizierung (siehe Abschnitt 4.3.3) der Suchbereich für Punktkorrespondenzen auf die gleiche Bildzeile reduziert werden kann. In [46] wird zur schnelleren Disparitätsberechnung eine nichtlineare Tiefenquantisierung durchgeführt, d.h. durch die Vorgabe von quantisierten Tiefenebenen (siehe Abbildung 6.2) ist die Zuordnung korrespondierender Punkte in zwei Ansichten nicht nur auf eine Linie sondern direkt auf die Auswahl einiger weniger Punkte beschränkt.

Allerdings besitzen SfS-Verfahren das Problem der Verdeckung. Eine Disparitätsschätzung von Punkten, die nur in einer Ansicht sichtbar sind und in weiteren Ansichten durch Vordergrundobjekte verdeckt werden, ist auf direktem Weg nicht möglich. Globale Optimierungsansätze wie z. B. *Markov Random Fields* (MRF) [92] oder partielle Differentialgleichungen (engl. *Partial Differential Equation*, PDE) [94, 169] versuchen indirekt die Tiefe



(a) Draufsicht (b) Vorderansicht

Abbildung 6.2.: Tiefenquantisierung zur Reduzierung des Suchbereichs korrespondierender Punkte

zu rekonstruieren. Ein guter Überblick und Vergleich von SfS-Verfahren ist in Scharstein und Szeliski [156] zu finden.

SfM wurde bereits in Kapitel 5 umfangreich beschrieben. Da zur Erzeugung dichter Tiefenkarten zeitlich versetzte Ansichten verwendet werden müssen, wird SfM fast ausschließlich auf statische Szenen angewendet, d.h. es können Disparitätsschätzverfahren wie bei SfS zum Einsatz kommen. Allerdings stellt die Rektifizierung, die bei SfS häufig verwendet wird, bei SfM ein größeres Problem dar, da die Kameraanordnung durch die Bewegung variiert und somit die Transformationsparameter ständig neu berechnet werden müssten. Demzufolge wird bei SfM häufig auf die Rektifizierung verzichtet.

Tiefe bzw. Form aus Silhouetten (engl. *Shape-from-Silhouette*, SfSi) ist zwar ein sehr leistungsstarkes Verfahren zur Rekonstruktion der Form von Objekten, benötigt allerdings für qualitativ hochwertige Ergebnisse eine große Zahl an Kameras, die das Objekt von allen Richtungen aus aufnehmen. Im Wesentlichen bestehen Objektsilhouetten aus binären Segmentierungsmasken, die die Szenen in Vordergrund (zur Silhouette zugehörig) oder Hintergrund aufteilen. Sind die Kameraparameter bekannt, kann die *visuelle Hülle* durch Rückprojektion der binären Segmentierungsmasken in den 3D-Raum über die Schnittmenge berechnet werden [20, 104]. SfSi-Verfahren werden vor allem bei *Free-Viewpoint Video* (FVV) [21, 197] eingesetzt und spielen bei Konvertierungsverfahren von Videoaufnahmen mit nur einer Kamera keine Rolle.

Ein guter Überblick über die vorgestellten Tiefenschätzverfahren ist in [34] zu finden. So verschieden diese Ansätze untereinander auch sind, besitzen sie eine Gemeinsamkeit. Sie unterliegen alle gewissen Einschränkungen und können nur bedingt eingesetzt werden. Sind z. B. keine unterschiedlichen Ansichten verfügbar - sei es, dass sich die Kamera

bei der Aufnahme nicht bewegt hat oder die Szene nicht mit mehreren Kameras synchron aufgezeichnet wurde - scheiden die Verfahren wie Tiefe aus Stereo, Bewegung und Silhouetten für die Tiefenschätzung aus. Besteht die aufgenommene Szene größtenteils aus homogenen Regionen, so ist eine Schätzung der Tiefe über die Textur fast ausgeschlossen. Ebenso die Disparitätsschätzung, wie sie überwiegend bei Tiefe aus Stereo und aus Bewegung zum Einsatz kommt, ist im Falle homogener Regionen nur bedingt möglich und äußerst fehlerbehaftet. Abhilfe schafft man sich hierbei u.a. durch die Rektifizierung, aber auch durch zusätzliche Tiefenquantisierung.

Ferner entstehen bei Kameraaufnahmen mit sehr kleinen Blenden und langen Belichtungszeiten kaum unscharfe Bereiche, so dass die Tiefenschätzung aus Fokus/Defokus praktisch nicht eingesetzt werden kann. Da bei vielen Filmaufnahmen gerade darauf geachtet wird, dass die aufgenommene Szene gut ausgeleuchtet ist, spielen ebenso Schattierungen eher eine untergeordnete Rolle.

Die aufgeführten Einschränkungen lassen sich natürlich durch den Einsatz hybrider Techniken (z. B. in [110, 155]), d.h. durch Kombination der Verfahren, reduzieren, allerdings sind damit nahezu perfekte dichte Tiefenkarten nicht gewährleistet, weshalb dieses Forschungsgebiet derzeit noch immer im aktuellen Fokus vieler Forschungseinrichtungen steht. Die Synthese neuer virtueller Ansichten aus Originalansichten und Tiefeninformation hängt allerdings qualitativ von den ermittelten Tiefenkarten ab.

6.1.3. Stereobildsynthese mittels DIBR

Sind Tiefeninformationen für jedes Pixel eines Bildes vorhanden, so können neue Ansichten prinzipiell über zwei Wege erzeugt werden. Zum einen kann jeder Punkt im Bild entsprechend seiner Tiefe in den 3D-Raum rückprojeziert und anschließend der erzeugte 3D-Punkt entsprechend Gl. 2.12 in die Bildebene der Kameras projeziert werden:

$$Z\mathbf{m} = P\mathbf{M} = K\left(R\tilde{\mathbf{M}} + \mathbf{t}\right) \tag{6.1}$$

$$Z'\mathbf{m}' = P'\mathbf{M} = K'\left(R'\tilde{\mathbf{M}} + \mathbf{t}'\right). \tag{6.2}$$

Abbildung 6.3 zeigt ein Beispiel für die Rückprojektion eines Punktes in den 3D-Raum und die anschließende Projektion des Punktes in die Bildebene einer virtuellen Kamera (rot). Sei Gl. 6.1 die Projektionsgleichung der Kamera mit der Originalansicht und Gl. 6.2 entsprechend die Projektionsgleichung für eine virtuelle Kamera, für die ein virtuelles Bild erstellt werden soll, und sei ferner das Weltkoordinatensystem im Ursprung der Originalkamera, dann ist $R = \underline{I}$ und $\mathbf{t} = (0,0,0)^T$. Da die virtuelle Kamera zur Erzeugung einer stereoskopischen Ansicht die gleichen internen Kameraparameter wie die Originalkamera besitzen sollte, kann K mit K' gleichgesetzt werden. Durch Umstellung von Gl. 6.1 ergibt sich

$$\tilde{\mathbf{M}} = ZK^{-1}\mathbf{m}. \tag{6.3}$$

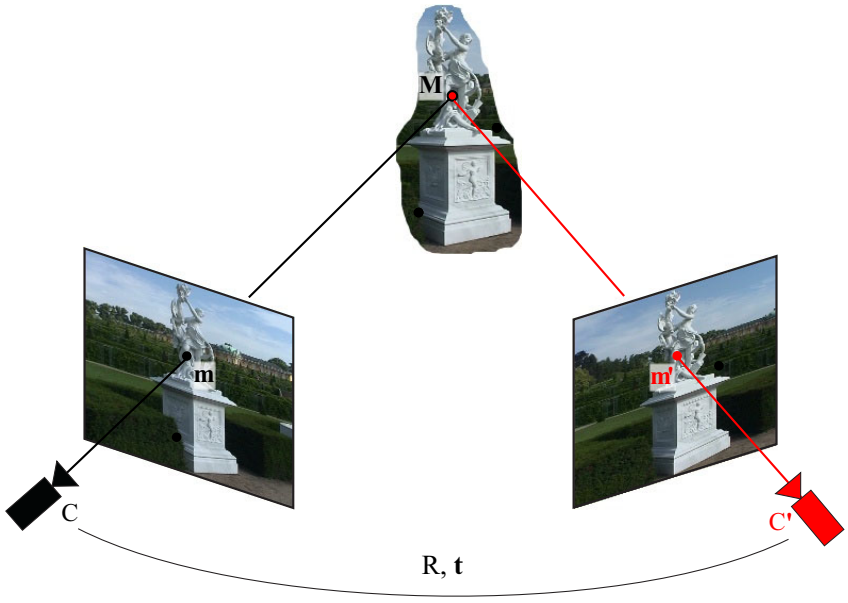


Abbildung 6.3.: Generierung virtueller Ansichten durch Projektion von 3D-Raumpunkten in die virtuelle Kamera

Setzt man Gl. 6.3 in Gl. 6.2 ein, erhält man

$$Z'\mathbf{m}' = ZKR'K^{-1}\mathbf{m} + K\mathbf{t}' \tag{6.4}$$

als *3D-Bildsynthesegleichung*, d.h. sind die internen und externen Parameter der virtuellen Kamera und die Tiefenwerte Z für alle Punkte bekannt, lässt sich die virtuelle Ansicht mit Gl. 6.4 erzeugen [44].

Allerdings muss darauf geachtet werden, dass Punkte, die eine größere Tiefe aufweisen, durch naheliegende Objekte bzw. Punkte in einer neuen Ansicht verdeckt werden können. Anders ausgedrückt können zwei unterschiedliche Punkte aus der Originalansicht auf ein und demselben Punkt bzw. Pixel im Pixelraster der virtuellen Ansicht abgebildet werden. In diesem Fall wird das sogenannte *Z-Buffering* verwendet, so dass Punkte, die dichter an der Kamera liegen, weiter entfernte Punkte in der Bildebene überschreiben. Eine einfache Methode zur Abarbeitungsreihenfolge der Punkte in der Originalansicht wäre demzufolge, mit den am weitesten entfernten Punkten zu beginnen und sukzessive alle weiteren Punkte bis zu den Punkten, die am dichtesten zur Kamera liegen, abzuarbeiten.

Die zweite Methode zur Erzeugung einer virtuellen stereoskopischen Ansicht aus einer Originalansicht mit Tiefenkarte lässt sich direkt aus den 2D-Bilddaten herleiten. Dazu sollen aber vorerst einige grundlegende Einschränkungen und Modifikationen erläutert werden, die den Einsatz dieses Verfahrens im Hinblick auf die Darstellung stereoskopischer Ansichten optimiert.

Shift-Sensor-Ansatz

Für die Erzeugung einer zweiten stereoskopischen Ansicht ist die Wahl eines geeigneten Kamerasetups erforderlich. Entsprechend dem menschlichen Augenabstand von durchschnittlich 64 mm (63 mm bei Frauen und 65 mm bei Männern) [108] kann die Position der virtuellen Kamera über die *interaxiale Distanz* t_x (Länge der Basislinie zwischen den Kameras) definiert werden. Die Augen konvergieren allerdings je nachdem, wo sich ein Objekt, welches fokusiert werden soll, im dreidimensionalen Raum befindet. Zwar lässt sich solch ein konvergentes Stereokamerasetup mit einer entsprechenden virtuellen Kamera generieren, allerdings erfolgt die Darstellung der beiden Ansichten auf einer Ebene (z. B. Display oder Leinwand), wodurch Bereiche, die vor bzw. hinter dem fokusierten Objekt liegen sowohl eine negative bzw. positive horizontale Parallaxe als auch eine leichte vertikale Parallaxe erhalten. Vertikale Parallaxe ist allerdings bei der Darstellung von stereoskopischen Filmen zu vermeiden, da diese eine erhöhte Augenbelastung verursacht [45,75,138].

Durch die Wahl eines parallelen Kamerasetups kann dieses Problem umgangen werden, allerdings entsteht dadurch der Effekt, dass lediglich negative horizontale Parallaxe auftritt, die umso größer wird, je dichter Objekte an der Kamera liegen. Für Objekte, die weit entfernt sind bzw. im Unendlichen liegen, ist die Parallaxe nahezu Null. Bei der Darstellung stereoskopischer Filme entsteht somit der Eindruck, dass die Szene komplett vor der Bildebene erscheint. Da aber Displays oder Leinwände eine physikalische Begrenzung (einen Rahmen) besitzen und Objekte seitlich aus dem Bildausschnitt verschwinden können und somit von dem Rahmen teilweise verdeckt werden, würde ein Konflikt der monokularen und stereoskopischen Tiefenwahrnehmung entstehen [109]. Durch Verlagerung der *Null-Parallaxen-Ebene* (engl. *Zero-Parallax-Setting*, ZPS) aus dem Unendlichen weiter in den Vordergrund lassen sich diese Konflikte beheben, da Teile der Szene durch die somit entstehende zusätzliche positive Parallaxe hinter der Bildebene erscheinen. Die Verlagerung der Null-Parallaxen-Ebene erfolgt über den Shift-Sensor-Ansatz [44,196], bei dem der Sensor der Kamera in horizontaler Richtung verschoben wird. Zwar ist die mechanische Änderung einer realen Kamera im Vergleich zur Wahl eines konvergenten Kamerasetups wesentlich aufwendiger, hingegen ist dies bei einer virtuellen Kamera weitaus einfacher. Abbildung 6.4 zeigt den Vergleich eines konvergenten und parallelen Kamerasetups mit dem Shift-Sensor-Ansatz, der lediglich - hier zur Veranschaulichung verstärkt dargestellt - auf die virtuelle Kamera (rot) angewendet wurde.

Die Verschiebung des Sensors der vituellen Kamera hat Einfluss auf die internen Kameraparameter und damit auf die Kamerakalibrierungsmatrix K. Sei K_S die Kalibrierungs-

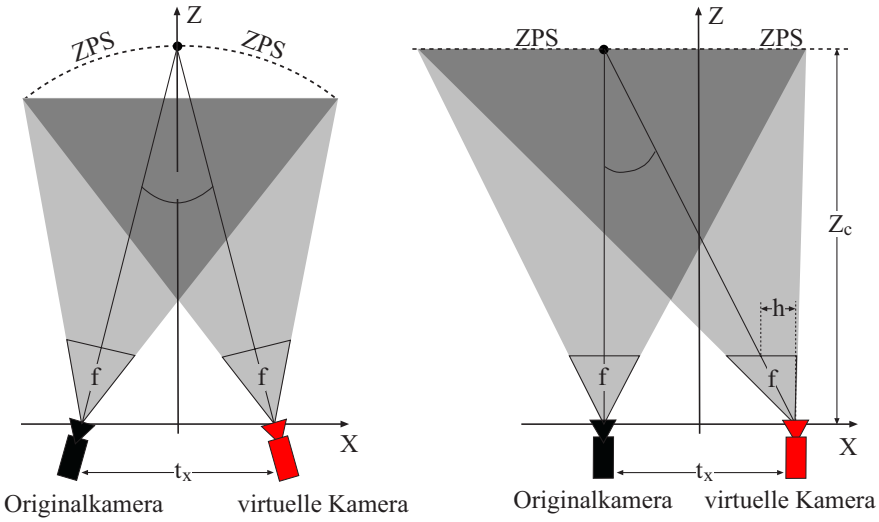


Abbildung 6.4.: Vergleich eines konvergenten und parallelen Kamerasetups mit Shift-Sensor

matrix der Kamera mit dem in horizontaler Richtung versetzten Sensor, so ergibt sich:

$$K_S = K + \begin{bmatrix} 0 & 0 & h \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{6.5}$$

Bei einem parallelen Kamerasetup und mit dem Weltkoordinatensystem im Ursprung der Originalkamera besitzen beide Kameras die gleichen Orientierungen. Somit ist $R = R' = \underline{I}$. Zudem ist die Tiefe Z und Z' eines 3D-Punktes von beiden Kamerazentren identisch. Mit

$$K_S R K^{-1} = K_S K^{-1} = \underline{I} + \begin{bmatrix} 0 & 0 & h \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{6.6}$$

lässt sich Gl. 6.4 zu

$$Z\mathbf{m}' = Z\left[\mathbf{m} + \begin{pmatrix} h \\ 0 \\ 0 \end{pmatrix}\right] + K_S\mathbf{t}' \tag{6.7}$$

bzw.

$$\mathbf{m}' = \mathbf{m} + \frac{K_S\mathbf{t}'}{Z} + \begin{pmatrix} h \\ 0 \\ 0 \end{pmatrix} \tag{6.8}$$

vereinfachen.

Da die Tiefeninformation in vielen Anwendungsgebieten nur in Form von 8-bit quantisierten Grauwerten (siehe Abbildung 6.1), die im Bezug auf eine vordere Tiefenebene Z_{nah} und eine hintere Tiefenebene Z_{fern} normiert sind, und nicht als metrisches Maß vorliegt, muss für Z eine Umrechnung dieser Grauwerte vorgenommen werden. Für die Umrechnung werden den beiden extremen Grauwerten die Tiefen Z_{nah} (Grauwert 255) und Z_{fern} (Grauwert 0) zugeordnet und mit

$$Z = Z_{fern} + \nu \frac{Z_{nah} - Z_{fern}}{255} \tag{6.9}$$

sämtliche Zwischenwerte berechnet, wobei ν der entsprechende Grauwert eines Pixels ist ($\nu = [0 \ \ldots \ 255]$).

Beispiel: Sei der Grauwert eines Pixels $\nu = 100$, die hintere Tiefenebene bei $Z_{fern} = 30\ m$ und die vordere Tiefenebene bei $Z_{nah} = 0\ m$, so ergibt sich für die metrische Tiefe des entsprechenden 3D-Punktes $Z = 30\ m - 100 \cdot (30\ m/255) = 18,24\ m$.

Da der Versatz der Pixel zwischen Originalansicht und stereoskopischer Ansicht lediglich horizontal ist, kann Gl. 6.8 weiter vereinfacht werden:

$$\begin{aligned} x' &= x + \Delta x \\ &= x + \frac{\alpha_x t_x}{Z} + h. \end{aligned} \tag{6.10}$$

Der Sensorversatz h wird entsprechend der gewählten Entfernung Z_C der Null-Parallaxen-Ebene und der interaxialen Distanz t_x berechnet:

$$h = -\frac{\alpha_x t_x}{Z_C}. \tag{6.11}$$

Die Parallaxe Δx eines Bildpunktes ergibt sich somit zu

$$\Delta x = \pm \alpha_x t_x \left(\frac{1}{Z} - \frac{1}{Z_C} \right), \tag{6.12}$$

wobei das Vorzeichen $\pm$ sowohl für die Synthese einer linken $(+)$ als auch einer rechten $(-)$ virtuellen Stereoansicht steht. Aus Gl. 6.12 wird deutlich, dass die Parallaxe neben der Tiefe eines Punktes durch das gewählte Kamerasetup variiert werden kann, indem die fokale Länge a_x (in Pixelkoordinaten, siehe Abschnitt 2.2.2), die interaxiale Distanz t_x bzw. die Tiefe der Null-Parallaxen-Ebene Z_C verändert werden.

Bei der Bildsynthese mit Gl. 6.10 treten drei grundlegende Probleme auf. Zum einen ist nicht gewährleistet, dass Pixel aus der Originalansicht durch den Versatz Δx einen ganzzahligen Wert in der virtuellen Ansicht erhalten und somit auf das quantisierte Pixelraster direkt abgebildet werden. Da aber der Versatz ausschließlich in horizontaler Richtung erfolgt, kann dieses Problem durch einfache lineare Interpolationsverfahren gelöst werden. Zum anderen entsteht das Problem der Sichtbarkeit einzelner Punkte oder Bereiche. Z. B.

können zwei oder mehr Punkte der Originalansicht aufgrund ihrer unterschiedlichen Tiefe und Lage im Bild auf dem selben Punkt in der virtuellen Ansicht abgebildet werden. Eine Lösung dieser Problematik stellt die Verwendung von Z-Buffern - also die zusätzliche Verarbeitung und Speicherung der Tiefenwerte - dar. Eine viel einfachere Methode wird in [123] vorgeschlagen, indem die Abarbeitungsreihenfolge der Pixel des Originalbildes sinnvoll festgelegt wird. Bei der Synthese einer linken virtuellen Ansicht müsste die Abarbeitungsreihenfolge spaltenweise vom rechten Rand des Originalbildes erfolgen, damit automatisch Vordergrundpixel die Hintergrundpixel überschreiben.

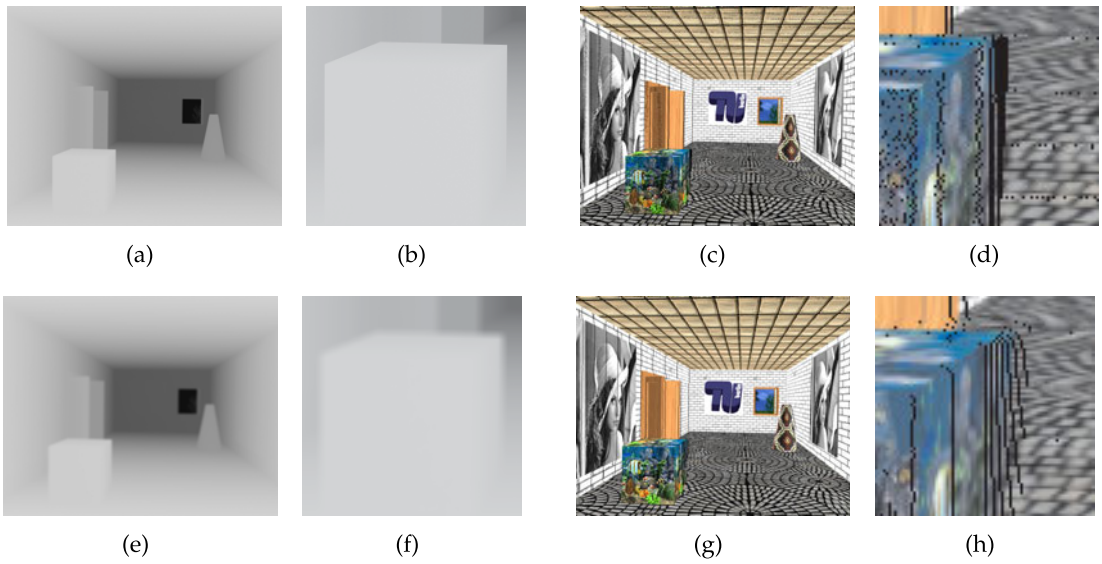
Das wohl größte Problem der Stereobildsynthese ist allerdings die Aufdeckungsproblematik, d.h. Teile des Hintergrunds werden in der virtuellen Ansicht sichtbar. Da diese Texturinformation allerdings weder im Originalbild noch in der Tiefenkarte enthalten ist, können Verfahren wie Interpolation, Extrapolation oder Spiegelung des Hintergrundes eingesetzt werden. Diese Verfahren verursachen je nach Größe des aufgedeckten Bereichs unterschiedliche Artefakte wie den sogenannten *rubber-sheet*-Effekt (engl., *Gummihauteffekt*), Auftreten von Streifen an den Objektkanten oder doppelte Konturen (Schattenbilder). Eine elegantere Lösung mit geringeren Auswirkungen auf die Bildqualität wird in [44] vorgestellt, indem eine Vorverarbeitung der Tiefenkarten vorgenommen wird, um Aufdeckungen und damit Löcher von vornherein zu vermeiden.

Vorverarbeitung der Tiefenkarten

Aufdeckungen bzw. Löcher entstehen vor allem an großen Tiefensprüngen, d.h. an den Außenkanten von Vordergrundobjekten, da der Übergang von einem nahen Vordergrundobjekt zum Hintergrund große Helligkeitsunterschiede der Grauwerte liefert und somit unmittelbar benachbarte Pixel verschieden stark in der virtuellen Ansicht versetzt werden. Abbildung 6.5(a) zeigt die Tiefenkarte und Abbildung 6.5(c) die erzeugte virtuelle Stereoansicht mit jeweils der Vergrößerung eines Bildausschnitts. Bei der Bildsynthese wurde zur Quantisierung bzw. Rundung der Pixelpositionen in der virtuellen Ansicht nur das *Nearest-Neighbor*-Verfahren angewendet, so dass einzelne Abtastlücken im Bild entstehen, welche aber durch bilineare Interpolation einfach gefüllt werden können. Die Aufdeckungen - z. B. am Würfel (siehe Vergrößerung in Abbildung 6.5(d) - stellen allerdings ein weitaus größeres Problem dar.

Zur Vermeidung von Aufdeckungen kann eine Filterung der Tiefenkarten angewendet werden, welche Tiefpasscharakteristik aufweisen muss, um hochfrequente bzw. „harte“ Übergänge an Objektkanten zu glätten. Fehn [44] verwendet als Tiefpass ein zweidimensionales Gaußfilter mit der Filterfunktion:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\left(x^2+y^2\right)/\left(2\sigma^2\right)}. \tag{6.13}$$



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

(e) (f) (g) (h)

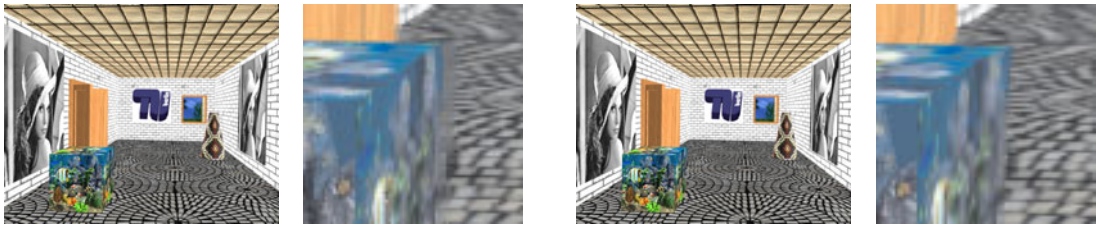
Abbildung 6.5.: Vorverarbeitung der Tiefenkarten durch Gaußfilterung: (a) ursprüngliche Tiefenkarte mit (b) Vergrößerung des Vordergrundobjekts, (c) generierte virtuelle Stereoansicht aus dem linken Originalbild mittels DIBR mit (d) Vergrößerung des Vordergrundobjekts, (e) tiefpassgefilterte Tiefenkarte mit (f) Vergrößerung des Vordergrundobjekts, (g) generierte virtuelle Stereoansicht aus dem linken Originalbild mittels Tiefpassfilterung und DIBR mit (h) Vergrößerung des Vordergrundobjekts. Einstellungen: $Z_C = 4\ m$, $t_x = 64\ mm$, $Z_{nah} = 0\ m$, $Z_{fern} = 20\ m$, $\alpha_x = 869,12\ Pixel$.

Sei $I_Z(x,y)$ die ursprüngliche Tiefenkarte, dann ergibt sich durch Filterung mit Gl. 6.13 die resultierende Tiefenkarte

$$J_Z(x,y) = I_Z(x,y) * G(x,y). \tag{6.14}$$

Abbildung 6.5(e) zeigt die tiefpassgefilterte Tiefenkarte und Abbildung 6.5(g) die erzeugte virtuelle Stereoansicht mit jeweiliger Vergrößerung eines Bildausschnitts. Bei der Bildsynthese wurde wiederum das Nearest-Neighbor-Verfahren verwendet, so dass durch Rundungen Abtastlücken entstehen. Die ohne Vorverarbeitung der Tiefenkarten entstandenen Aufdeckungen sind hier nicht mehr vorhanden. In horizontaler Richtung treten lediglich einzelne Löcher der Breite eines Pixels auf.

Anstelle der Nearest-Neighbor-Methode sollte vor allem bei parallelen Kamerasetups eher eine lineare Interpolation verwendet werden, da diese für einen rein horizontalen Versatz der Pixel einfach zu realisieren ist [45]. In Abbildung 6.6 sind die vollständig synthetisierten virtuellen Stereoansichten dargestellt. Die Abtastlücken treten durch die lineare Interpolation nicht mehr auf. Bei Abbildung 6.6(a) wurde die virtuelle Ansicht aus dem Originalbild und der Tiefenkarte ohne Tiefpassfilterung erstellt. Die aufgedeckten Bereiche wurden durch Hintergrundextrapolation mit Texturinformation aufgefüllt (siehe Vergrößerung in Abbildung 6.6(b)). Abbildung 6.6(c) zeigt die virtuelle Ansicht



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

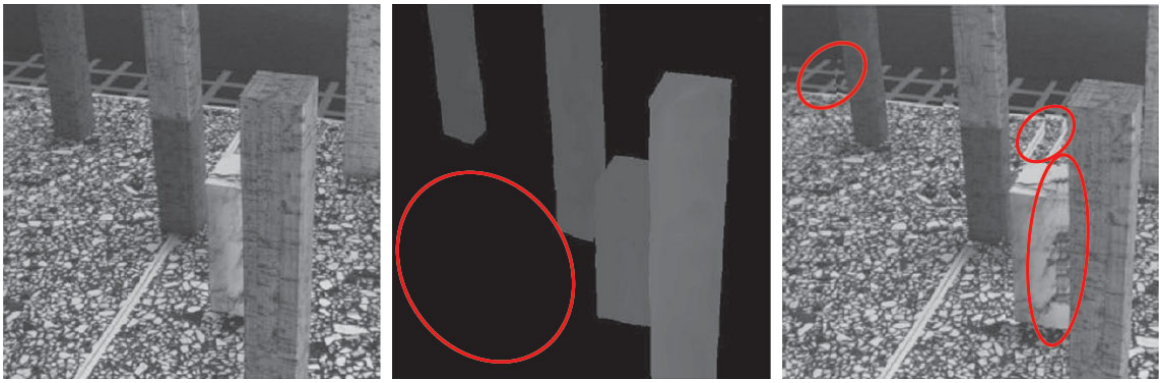
Abbildung 6.6.: Stereobildsynthese mit (a)-(b) Hintergrundextrapolation bzw. mit (c)-(d) Gaußfilterung der Tiefenkarten

bei Verwendung einer Gaußfilterung der Tiefenkarte. Da keine Aufdeckungen vorhanden sind, wurde ausschließlich linear interpoliert, um Abtastlücken zu vermeiden. Durch die Gaußfilterung werden aber auch die Kanten des Vordergrundobjekts leicht mit dem Hintergrund „verschmiert“, d.h. nicht nur die Interpolation sondern auch die Filterung der Tiefenkarten bewirkt einen Tiefpasseffekt in der virtuellen Ansicht (siehe Vergrößerung in Abbildung 6.6(d)). Allerdings scheinen diese Artefakte subjektiv geringere Auswirkungen auf die Tiefenwahrnehmung bei Betrachtung des Stereobildpaares zu haben [45].

6.1.4. Aktuelle 2D/3D-Konvertierungsverfahren basierend auf DIBR

Forschungsarbeiten, die sich gesamten Systemen zur Konvertierung von monokularen Videosequenzen in stereoskopische Sequenzen widmen, sind überwiegend erst in den letzten Jahren in den Fokus gerückt. Die meisten dieser Verfahren berechnen, wie zuvor beschrieben, erst die Tiefenkarten für jedes Bild einer Videosequenz, um anschließend neue stereoskopische Ansichten mittels DIBR zu generieren [31, 63, 93, 122, 129]. Die Rekonstruktionsergebnisse zeigen allerdings, dass die aufgeführten Probleme bei der Bestimmung dichter Tiefenkarten große Fehler verursachen, die in die folgende Synthese mittels DIBR einfließen.

Abbildung 6.7 zeigt die ermittelten Tiefenkarten und die rekostruierten Stereoansichten aus [129] mit einigen rot markierten Fehlern. Bei den Tiefenkarten sind lediglich für einzelne Objekte die Tiefen ermittelt worden. Z. B. hat der Boden in Abbildung 6.7(a) laut Tiefenkarte eine konstante Tiefe. Bei den synthetisierten Stereoansichten ist deutlich die Aufdeckungsproblematik zu erkennen (siehe rot umrandete Bereiche). Vor allem die Fehler an den gradlinigen Kanten in Abbildung 6.7(a) haben einen hohen störenden Einfluss beim räumlichen Sehen. In Abbildung 6.8 sind die Ergebnisse (Tiefenkarten) des 2D/3D-Konvertierungssystems nach [93] dargestellt. Einige der falsch ermittelten Tiefen sind wiederum rot markiert. Fehler, wie sie z. B. an dem „Rettungsring“ im dritten Bild von links auftreten (der Ring hat aufgrund unterschiedlicher Farbe verschiedene Tiefen) bzw. die Oberkörper der Personen im rechten Bild haben eine stark abweichende Tiefe von den



(a)

(b)

Abbildung 6.7.: Ergebnisse der 2D/3D-Konvertierung nach [129]. Links: Originalansicht, mitte: erzeugte Tiefenkarte, rechts: virtuelle Stereoansicht. (Quelle: [129])

Abbildung 6.8.: Ergebnisse der 2D/3D-Konvertierung nach [93]. Oben: Originalansicht, unten: erzeugte Tiefenkarten. (Quelle: [93])

Beinen, verursachen neben den Fehlern, die zusätzlich bei der Stereobildsynthese mittels DIBR entstehen, starke Irritationen bei der Tiefenwahrnehmung.

Da die automatische Erzeugung von Tiefenkarten noch sehr begrenzt gute Ergebnisse liefert, setzten einzelne Unternehmen wie z. B. *DDD*[1] auf semiautomatische bzw. manuelle Konvertierungsverfahren. In [63] wird ein interaktives System vorgestellt, bei dem

[1] Dynamic Digital Depth

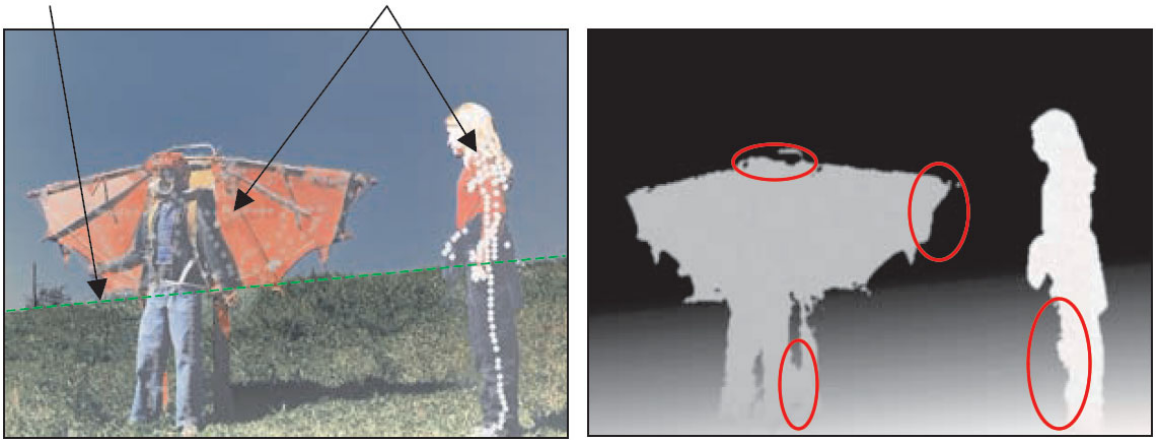


Abbildung 6.9.: Ergebnisse der 2D/3D-Konvertierung nach [63]. Links: Originalansicht, rechts: semiautomatisch erzeugte Tiefenkarte. (Quelle: [63])

Bildbereiche in einzelnen Schlüsselbildern markiert und diesen manuell Tiefen zugewiesen werden. Zwischen den Schlüsselbildern wird die Tiefe interpoliert. In Abbildung 6.9 sind die Ergebnisse (Tiefenkarte) dieses Ansatzes für ein Schlüsselbild dargestellt. Die linke Abbildung zeigt die Originalansicht mit einzelnen markierten Bereichen (Horizont, Personen) und die rechte Abbildung die erzeugte Tiefenkarte. Einige der Tiefenfehler sind wieder rot markiert.

6.2. Realistisches Stereo durch 2-D Ebenentransformation

Die größte Problematik bei DIBR-Ansätzen ist die Berechnung einer dichten Tiefenkarte für jedes Bild einer Videosequenz. Bislang wird die Bestimmung von dichten Tiefenkarten immer noch als „nicht vollständig lösbar“ bezeichnet [3,168]. Vor allem Sequenzen mit vielen homogenen Regionen sind für Disparitätsschätzungen denkbar ungeeignet und sehr fehleranfällig. Eine weitere Problematik ist die zeitliche Konsistenz der Tiefeninformation. Die relative Tiefe der aufgenommenen Szene kann durch die Disparitäten dargestellt werden. Große Disparitäten weisen zum Beispiel auf ein Vordergrundobjekt hin. Allerdings können sich diese Disparitäten nicht nur durch die Variation der Tiefe im Laufe der Sequenz ändern, sondern auch durch eine Variation der Kameraführung. Zum Beispiel kann die Geschwindigkeit oder die Orientierung der Kamera stark variieren. Die Folge ist dann eine starke Schwankung der Disparitäten, obwohl sich die Tiefe der Objekte ggf. gar nicht geändert hat.

Speziell für die Synthese stereoskopischer Ansichten aus monokularen Videosequenzen existieren noch weitere Ansätze, die ohne 3D-Geometrieinformation auskommen. Zu ihnen zählen Verfahren, die z. B. eine räumliche und zeitliche Interpolation bewegter Bilder durchführen [55] bzw. eine planare Transformation auf zeitlich benachbarte Bilder

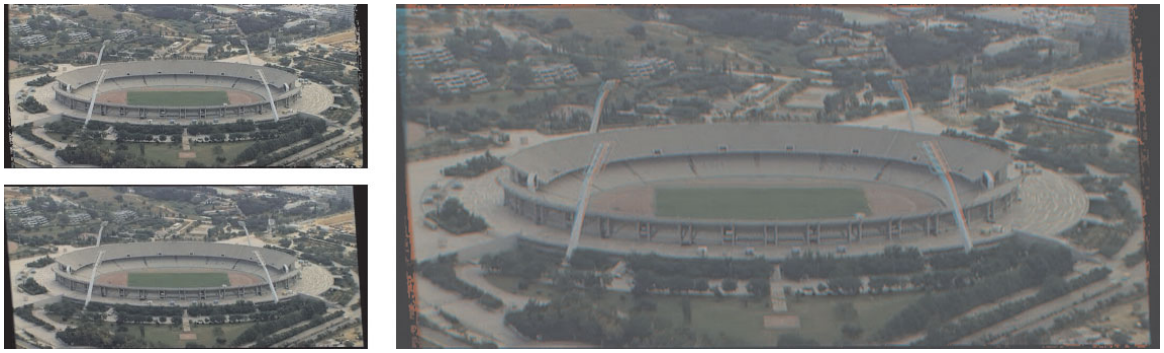


Abbildung 6.10.: Ergebnisse der 2D/3D-Konvertierung nach [152]. Links: erzeugte Stereoansichten, rechts: überlagertes Stereobild (anaglyph). (Quelle: Rafael Inc., Israel)

anwenden [152], um somit ein Stereokamerasetup zu simulieren. Allerdings ist bei diesen Ansätzen die zeitliche Konsistenz der Parallaxe über die Bildsequenz nicht gegeben und hängt stark von der 3D-Szene und der Kamerabewegung ab, da ein Stereokamerasetup nicht korrekt modelliert wird. Abbildung 6.10 zeigt ein einzelnes Bildpaar aus der erzeugten stereoskopischen Videosequenz nach [152]. Die Transformation zweier zeitlich versetzter Bilder auf eine gemeinsame Ebene erfolgt über die dominante Ebene in der Szene (hier die Erdoberfläche). Die Parallaxe, die den Tiefeneindruck vermittelt, nimmt mit größerem Abstand zur dominanten Ebene zu und nicht mit der eigentlichen Tiefe der Szene. Dieser Fehler ist besonders gut bei den Flutlichtmasten zu beobachten, bei denen die Parallaxe nach oben hin zunimmt.

In diesem Abschnitt wird ein neues Verfahren vorgestellt, das ohne dichte Tiefenkarten stereoskopische Ansichten erstellt. Als Basis dient lediglich die Information über den Kamerapfad und einige wenige 3D-Punkte, um neue virtuelle photokonsistente Ansichten mit einem möglichst realistischen 3D-Tiefeneffekt (engl. *Realistic Stereo View Synthesis*, RSVS) zu erzeugen [101, 102].

6.2.1. Problemstellung und Überblick über die Methode

Der neue Ansatz zur Generierung virtueller stereoskopischer Ansichten basiert auf zwei grundlegenden Verfahren der Computer Vision, der 3D-Rekonstruktion aus mehreren Ansichten (SfM) [85] und dem *bildbasierten Rendern* (engl. *image based rendering*, IBR) [123]. Im ersten Schritt werden die Kameraparameter und die grobe 3D-Struktur der aufgenommenen Szene in Form einzelner 3D-Punkte $\mathbf{M}_i$ geschätzt. Die Rekonstruktion aus mehreren Ansichten und deren Einschränkungen sind bereits in Kapitel 5 ausgiebig behandelt worden. Anschließend wird für jede Kameraposition des geschätzten Kamerapfads eine virtuelle Stereokamera definiert. Die Bildinformation der virtuellen Kameras wird letztendlich

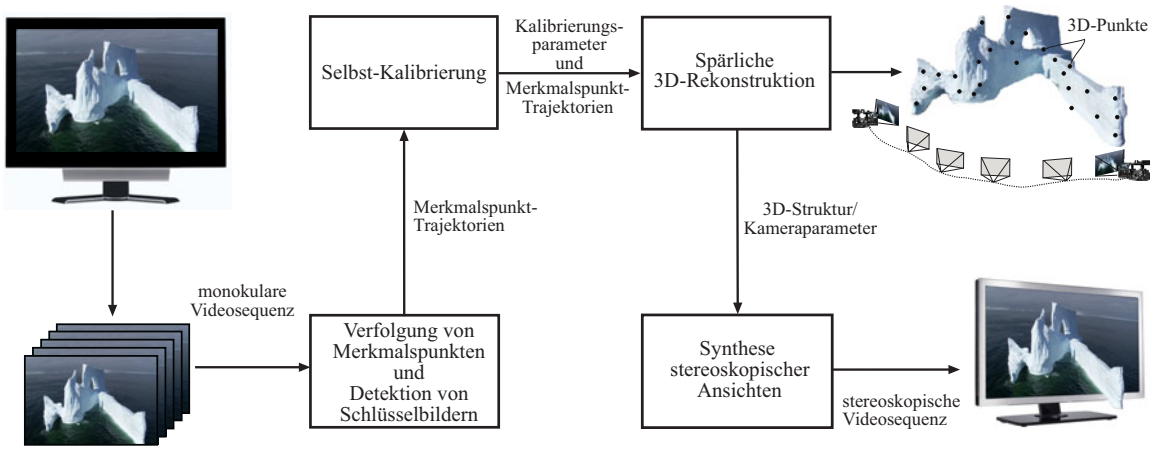


Abbildung 6.11.: Überblick über das 2D/3D-Konvertierungssystem.

über 2D-Homographien (perspektivische Transformationen) aus den umliegenden Originalkameras gewonnen, indem grundlegende Ansätze von IBR angewendet werden. Abbildung 6.11 zeigt noch einmal einen Überblick über den gesamten Konvertierungsprozess von 2D nach stereoskopisch 3D.

Die Probleme, die dieses Verfahren mit sich führt, liegen vor allem in dem Einsatz einer 2D-Transformation, die für die Synthese einer virtuellen Ansicht verwendet wird, da diese nur eine Approximation der eigentlichen Abbildung der 3D-Szene in der virtuellen Kamera darstellt. Die bei der Synthese entstehenden Transformationsfehler werden umso größer, je weiter sich die virtuelle Kamera von dem Kamerapfad entfernt bzw. je größer die Länge der Basislinie (siehe Abschnitt 5.3) zwischen virtueller Kamera und für die Transformation verwendeter Originalkamera wird. In Abbildung 6.12 ist ein Beispiel für den RSVS-Ansatz dargestellt. Der Originalkamerapfad und die rekonstruierten 3D-Punkte wurden mit SfM berechnet. Für jede Kameraposition (hier nur für ein paar ausgewählte Originalkameras angedeutet) wird mindestens eine parallele virtuelle Kamera (in diesem Fall acht virtuelle Kameras, deren Ansichten für spezielle autostereoskopische 3D-Bildschirme notwendig sind) definiert.

Die Abbildung zeigt, dass bei horizontaler Bewegung der Kamera die virtuellen Kameras sehr nah am Kamerapfad liegen und somit eine 2D-Transformation als Approximation der eigentlichen 3D-Transformation gewährleistet ist. Allerdings wird in der Vergrößerung unten links schon angedeutet, dass der Abstand zwischen Kamerapfad und den äußeren virtuellen Kameras größer wird, wenn die Kamerabewegung z. B. vorwärts gerichtet ist. Ein großer Vorteil von RSVS gegenüber DIBR ist aber die Tatsache, dass Hintergrundbereiche durch die Kamerabewegung aufgedeckt und somit tatsächlich in der virtuellen Ansicht sichtbar werden. Durch den Einsatz der 2D-Transformation wird die aufwendige und meist ungenaue Berechnung dichter Tiefenkarten umgangen.

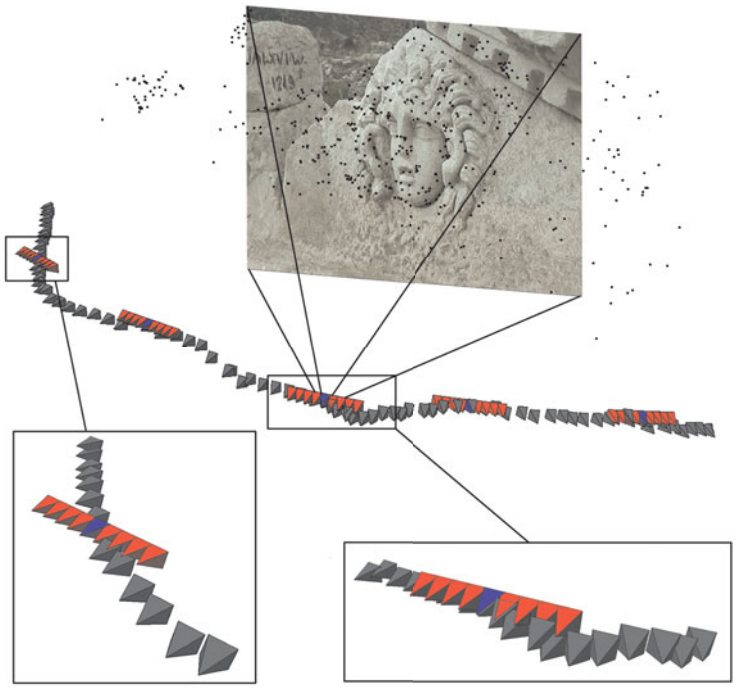


Abbildung 6.12.: Beispiel für RSVS: 3D-Rekonstruktion mit geschätztem Kamerapfad (grau) und virtuellem Stereo- bzw. Multiview-Kamerasetup (rot) für ein paar ausgewählte Originalkameras (blau)

6.2.2. Definition virtueller Kameras

Der erste Schritt zur Generierung neuer stereoskopischer Ansichten ist die Platzierung der virtuellen Kameras für jede Originalansicht der Videosequenz in die rekonstruierte 3D-Szene. Da SfM lediglich Rekonstruktionsergebnisse mit relativen Größenverhältnissen liefert, d.h. die 3D-Szene nur in Abhängigkeit eines Skalierungsfaktors berechnet werden kann, sind absolute Abstände von Objekten oder zwischen Kamerapositionen in Metern nicht vorhanden. Um eine virtuelle Stereokamera in den Kamerapfad zu integrieren, müssen die absoluten Entfernungen allerdings bekannt sein bzw. vorgegeben werden. Dabei ist es nicht unbedingt erforderlich bzw. jederzeit wünschenswert, die tatsächlichen absoluten Größen zu bestimmen, da ggf. Miniaturmodelle aufgenommen wurden, die aber eine andere Wirkung auf den Betrachter erzielen sollen, wie in Abbildung 6.13 beispielhaft dargestellt wird. Der Delorean aus *Zurück in die Zukunft* (engl. *Back to the Future*) ist hier nur ein Modellauto mit etwa 1/6 der Originalgröße. Das Bild soll aber den Eindruck eines Originals vermitteln. Der 3D-Effekt mit der entsprechenden Parallaxe sollte demnach über ein virtuelles Stereokamerasetup erfolgen, welches weiterhin das eigentliche Erscheinungsbild suggerieren soll.

Um eine gemeinsame Basis der relativen 3D-Struktur und des menschlichen Augenabstands zu finden, ist eine Interaktion bzw. ein Initialisierungsprozess erforderlich, d.h.

(a) (b)

Abbildung 6.13.: Filmauschnitt aus *Zurück in die Zukunft* (Quelle: Universal Studios, www.bttfmovie.com)

vor der Synthese muss ein absolutes Distanzmaß in Metern vorgegeben werden. Da die einzelnen 3D-Punkte und der Kamerapfad bis auf einen Skalierungsfaktor geschätzt wurden, könnte z. B. die Entfernung einer dieser Punkte zu einer Kameraposition innerhalb des Kamerapfads vorgegeben werden. Ohne Beschränkung der Allgemeinheit wird in den folgenden Berechnungsschritten der Abstand t_s des Schwerpunktes der 3D-Struktur zur ersten Kameraposition der Filmsequenz als Initialisierung verwendet. Liegt das Weltkoordinatensystem im Schwerpunkt der 3D-Struktur, lässt sich die absolute Position aller Kameras mit

$$C_j^m = t_s \frac{C_j}{\|C_1\|} \tag{6.15}$$

bestimmen, wobei $\|C_1\|$ die Vektornorm der ersten Kamera ist. Superskript m zeigt die Normierung in Metern an. Die Positionen der entprechenden virtuellen Stereokameras sind dann

$$C_{stereo,j}^m = C_j^m + R_j^{-1} \cdot [\pm t_x \quad 0 \quad 0]^T , \tag{6.16}$$

wobei $t_x = 64\ mm$ der durchschnittliche menschliche Augenabstand ist.

Da die internen Kameraparameter bis auf den Shift-Sensor und die Rotationsmatrizen der virtuellen Kameras bei der Wahl eines parallelen Kamerasetups identisch mit denen der Originalkameras sein sollen, können die Projektionsmatrizen der virtuellen Kameras nach Gl. 2.15 wie folgt berechnet werden:

$$P_{stereo,j}^m = K_{S,j} R_j \left[\underline{I} \mid \widetilde{C}_{stereo,j}^m \right] . \tag{6.17}$$

Die ermittelten 3D-Punkte $\mathbf{M}_i$ lassen sich über die Projektionsmatrizen $P_{stereo,j}^m$ mit

$$\mathbf{m}_{stereo,i,j} = P_{stereo,j}^m \, \mathbf{M}_i \tag{6.18}$$

in die Bildebene der virtuellen Kameras bzw. mit

$$\mathbf{m}_{i,j} = P_j^m \, \mathbf{M}_i \tag{6.19}$$

in die Bildebene der Originalkameras projizieren[2], so dass 2D-Punktkorrespondenzen $\{\mathbf{m}_{i,j} \leftrightarrow \mathbf{m}_{stereo,i,j}\}$ zwischen den Original- und virtuellen Ansichten gebildet werden können.

6.2.3. Bestimmung der 2D-Homographien

Für geringe Abstände - d.h. kleine Basislinienlänge - zwischen Original- und virtuellen Kameras lassen sich die 2D-Punktkorrespondenzen durch eine Homographie H in Form einer Ebenentransformation - z. B. die perspektivische Transformation - in Beziehung setzen:

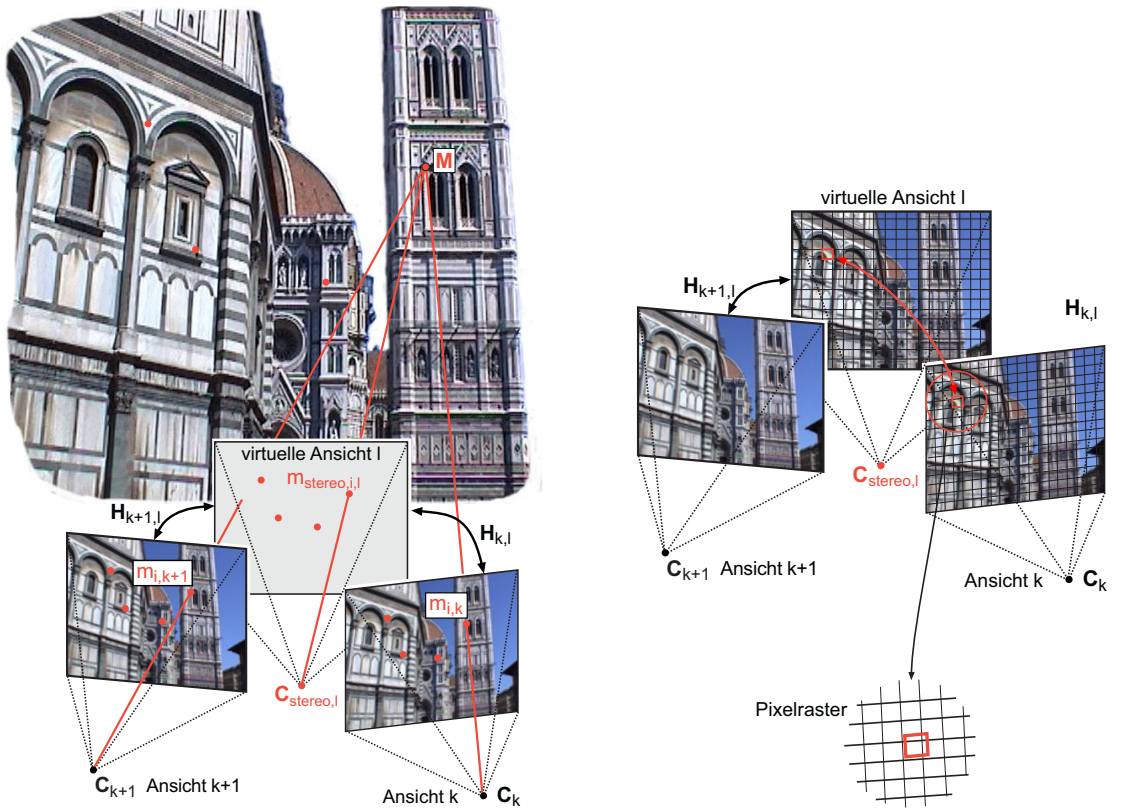
$$\mathbf{m}_{stereo,i,l} = H_{k,l} \, \mathbf{m}_{i,k} \qquad mit \; k \neq l \; . \tag{6.20}$$

Die Indizes k und l stehen für zwei beliebige nah beieinander liegende Kameras. $H_{k,l}$ ist eine 3×3-Matrix mit acht Freiheitsgraden und lässt sich allgemein mit mindestens vier Punktkorrespondenzen, die jeweils zwei unabhängige Gleichungen liefern (Gl. 2.33 und Gl. 2.34) und von denen jeweils drei Punkte nicht kollinear sein müssen, über einen DLT-Algorithmus oder nichtlinearen Ansatz bestimmen [68]. Die Bestimmung der Homographie mittels robuster Ansätze wird im Allgemeinen bevorzugt, um schlechte Punktkorrespondenzen zu eliminieren (siehe Kapitel 3). In diesem Fall ist allerdings von robusten Ansätzen abzuraten, da zum einen bereits robuste Verfahren bei der Berechnung der 3D-Struktur verwendet wurden und somit die Zuverlässigkeit der Punktkorrespondenzen gegeben ist und zum anderen robuste Ansätze zur Homographieberechnung verstärkt Punktkorrespondenzen eliminieren würden, die nicht auf einer dominanten Ebene liegen. Zwar wären die Ergebnisse in den Bereichen der Ansicht, in denen die dominante Ebene sichtbar ist, besser als bei einem Verzicht robuster Verfahren, allerdings würden dadurch verstärkt Transformationsfehler in den restlichen Bildbereichen auftreten.

Abbildung 6.14(a) stellt den Ansatz von RSVS grafisch dar, indem die 3D-Punkte - hier nur angedeutet mit vier 3D-Punkten, zwei Originalkameras und einer virtuellen Kamera - in die Bildebenen der Kameras projeziert werden und somit als 2D-Punktkorrespondenzen zur Bestimmung der Homographie zwischen einer Original- und einer virtuellen Kamera zur Verfügung stehen.

Die Berechnung der Homographien zwischen Ansichten, die von unterschiedlichen Positionen aufgenommen wurden, ist generell eine kritische Aufgabe, da für große Abstände der Kamerazentren erhebliche Transformationfehler in Form von perspektivischen

[2]Die Projektion der 3D-Punkte in die Originalansichten ist für die weiteren Betrachtungen nur notwendig, wenn die 2D-Punkte von der Merkmalspunktverfolgung (siehe Abschnitt 5.2) nicht vorhanden sind.



(a) Projektion und Bestimmung der 2D-Homographie

(b) Prinzip der Synthese einer virtuellen Ansicht

Abbildung 6.14.: Bestimmung der 2D-Homographie und Stereobildsynthese

Verzerrungen auftreten können, wenn die abgebildete Szene nicht aus nur einer Ebene besteht. Da aber die virtuelle Stereoansicht gerade für räumliche Szenen angewendet werden soll, um einen entsprechenden realistischen Tiefeneindruck zu erzeugen, sind diese Fehler möglichst gering zu halten, d.h. es werden für die Transformation nur Originalansichten verwendet, die dicht an der virtuellen Kamera liegen.

6.2.4. Stereobildsynthese

Die Synthese bzw. das „Auffüllen“ einer virtuellen Ansicht mit Bildinformationen kann über die in Gl. 6.20 ermittelte Homographie zwischen der virtuellen Ansicht und der am dichtesten liegenden Kamera des Originalkamerapfads erfolgen. Allerdings werden bei der Transformation Pixel von der Originalansicht wieder nicht direkt auf dem Pixelraster der virtuellen Ansicht abgebildet. Durch die Anwendung der Nearest-Neighbor-Methode würden somit Abtastlücken entstehen. Eine Möglichkeit besteht wiederum durch Interpolationsverfahren, allerdings sind diese komplexer als bei DIBR, da der Versatz der Pixel nicht zwingend in der gleichen Bildzeile erfolgt. Durch Inversion des Syntheseprozesses,

d.h. durch Inversion der Homographie H in Gl. 6.20, können die Texturinformationen für jedes Pixel der virtuellen Ansicht aus der Originalansicht transformiert werden:

$$\mathbf{m}_{i,k} = H_{k,l}^{-1}\, \mathbf{m}_{stereo,i,l} \ . \tag{6.21}$$

Sei $\mathcal{T}$ die Funktion, die die Parameter der perspektivischen Transformation beinhaltet, so lässt sich der zu ermittelnde Farbwert des Pixels (x, y) in der virtuellen Ansicht $I_{stereo,l}(x, y)$ mit

$$I_{stereo,l}\,(u, v) = I_k\,(\mathcal{T}(u, v)) \tag{6.22}$$

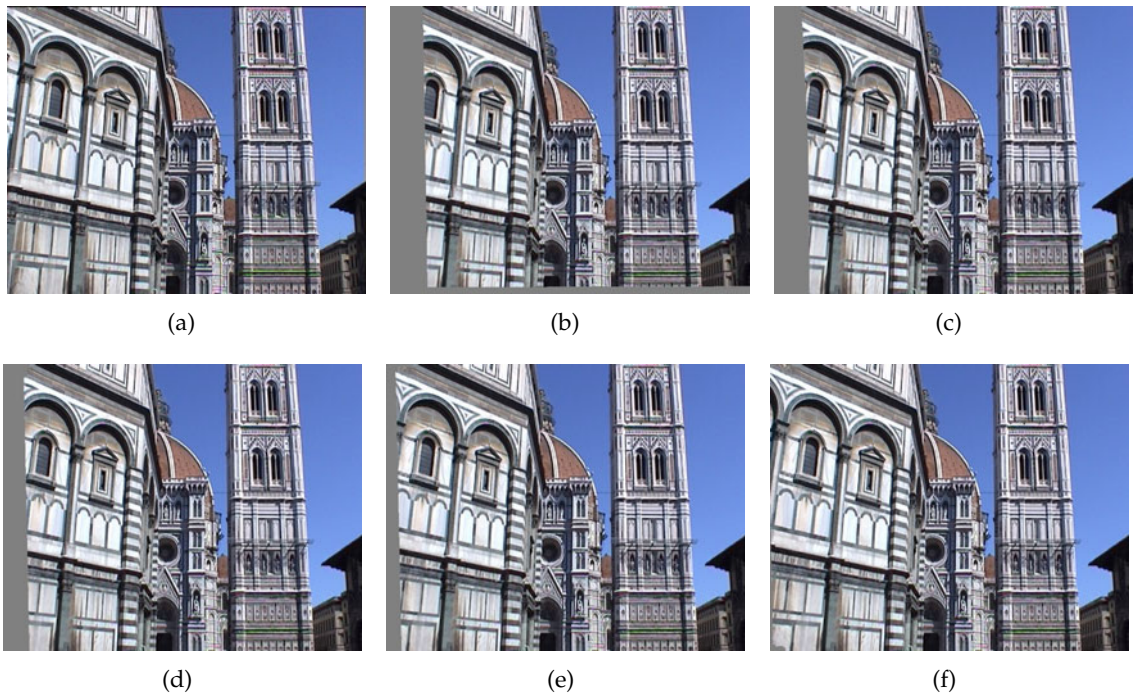
ermitteln. Allerdings liegt $\mathcal{T}(x, y)$ auch hier nicht direkt auf dem Pixelraster der Originalansicht. Durch bilineare oder bikubische Interpolation kann der Farbwert des Subpixels aus den umliegenden Pixeln durch Gewichtung der Farbinformation berechnet werden (siehe Abschnitt 5.2.4).

In Abbildung 6.14(b) ist das Prinzip der Synthese einer virtuellen Ansicht aus einer naheliegenden Originalansicht anschaulich dargestellt. Die Abbildung zeigt zudem, dass die Orientierung der Originalkamera zum Teil stark von der Orientierung der virtuellen Kamera abweichen kann, d.h. die Bildinformation aus der Originalansicht reicht nicht aus, um die virtuelle Ansicht komplett mit Bildinhalten zu füllen. Bei der Orientierung der beiden Kameras würden in diesem Fall am rechten Bildrand keinerlei Bildinformationen zur Verfügung stehen, da die virtuelle Kamera durch eine Rechtsrotation neue Bereiche der aufgenommenen Szene auf der rechten Seite aufdeckt.

Abbildung 6.15 stellt beispielhaft diese Problematik dar. In Abbildung 6.15(a) wird ein Originalbild der handgeführten Videosequenz *Dom*, zu dem ein virtuelles Stereobild erstellt werden soll, gezeigt. Das entsprechende virtuelle Stereobild, welches nur aus der am dichtesten liegenden Ansicht des Originalkamerapfads erstellt wurde[3], ist in Abbildung 6.15(b) dargestellt. In diesem Fall war die Orientierung der virtuellen Kamera gegenüber der dichtesten Originalkamera leicht nach unten links gerichtet, wodurch aufgedeckte Bereiche am linken und unteren Bildrand entstehen.

Um diese aufgedeckten Bereiche mit Bildinformationen zu füllen, müssen zusätzliche umliegende Kameraansichten hinzugezogen werden. In Abbildung 6.14(b) deckt z. B. Ansicht $k+1$ Teile der Szene auf der rechten Seite auf. Für die zusätzlich verwendeten Originalansichten müssen ebenfalls die Homographien zur virtuellen Ansicht berechnet und die Farbwerte entsprechend Gl. 6.22 in die Bildebene der virtuellen Kamera transformiert werden. Die Abbildungen 6.15(c)-6.15(f) zeigen die Ergebnisse des Auffüllens der virtuellen Stereoansicht mit bis zu 62 umliegenden Originalansichten.

[3] Es sei darauf hingewiesen, dass das Originalbild, **mit** dem das virtuelle Stereobild erstellt wird, nicht identisch mit dem Originalbild ist, **für** das ein Stereobild erstellt werden soll.



(a) (b) (c)

(d) (e) (f)

Abbildung 6.15.: Stereosynthese mit umliegenden Originalansichten: (a) linke Originalansicht, rechte virtuelle Stereoansicht erstellt aus (b) der dichtesten Originalansicht, (c) 10, (d) 25, (e) 40 und (f) 62 umliegenden Originalansichten.

Die Synthese einer virtuellen Stereoansicht aus benachbarten Originalansichten ist somit ein iterativer Prozess, der stark von der Kameraführung abhängt. Da es nicht immer möglich ist, die virtuelle Ansicht vollständig mit Bildinformationen aus den benachbarten Ansichten zu füllen, müssen Abbruchkriterien definiert werden, die diesen Prozess z. B. im Falle eines zu großen Transformationsfehlers beenden.

Die erste Abbruchbedingung ist der bei der Homographiebestimmung entstehende Transformationsfehler, d.h. der euklidische Abstand zwischen den Punkten $\mathbf{m}_{stereo,i,l}$ und den über die Homographie transferierten Punkten $H_{k,l}\ \mathbf{m}_{i,k}$. Ist der Medianwert der Transformationsfehler ϵ_i

$$\epsilon = \underset{i}{\text{median}}\ \epsilon_i, \quad \epsilon_i = d\left(\mathbf{m}_{stereo,i,l}, H_{k,l}\mathbf{m}_{i,k}\right) \tag{6.23}$$

größer als ein Schwellwert (z. B. 0,6 Pixel), werden keine weiteren Originalansichten für die Stereobildsynthese verwendet.

Als zweite Abbruchbedingung dient der Grad der Rekonstruktion η einer virtuellen Ansicht. Sind mehr als 99,5 Prozent des virtuellen Bildes rekonstruiert ($\eta = 0,995$), wird die Stereobildsynthese aus umliegenden Ansichten beendet. Da die noch fehlende Farbinformation lediglich im Randbereich der Bilder liegt, kann diese durch Extrapolation oder Spiegelung benachbarter Farbwerte aufgefüllt werden. In Abbildung 6.16 sind die

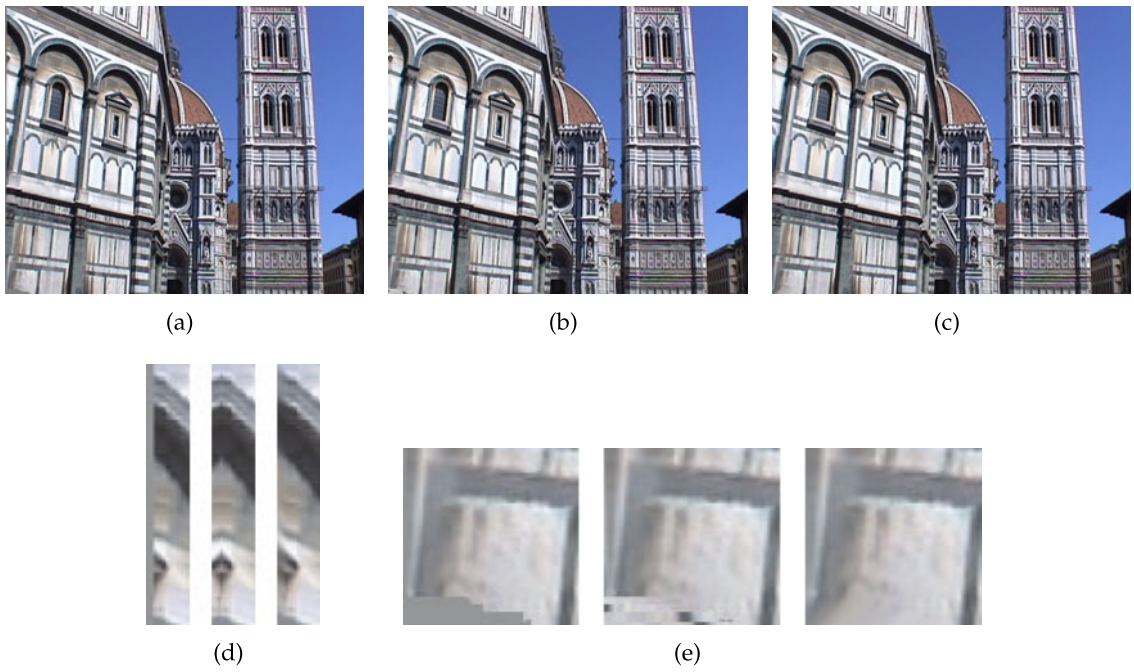


Abbildung 6.16.: Auffüllen des Randbereichs: (a) mittlerer Grauwert, (b) Spiegelung und (c) Extrapolation der Farbwerte benachbarter Pixel, (d) und (e) Vergrößerung der Bildausschnitte im Randbereich

Verfahren der Fehlerverschleierung im Randbereich des virtuellen Stereobilds dargestellt.

Die einfachste Fehlerverschleierung durch Verwendung des mittleren Grauwerts zeigt Abbildung 6.16(a). Die Spiegelung in Abbildung 6.16(b) weist stark sichtbare Fehler an Kanten und Linien auf (siehe jeweils die mittlere Vergrößerung in Abbildung 6.16(d) und 6.16(e)), bei denen der ursprüngliche Verlauf nicht mehr gegeben ist. Bei der Extrapolation in Abbildung 6.16(c) sind die Fehler visuell am linken Bildrand nicht mehr zu erkennen. Lediglich in der unteren linken Ecke des Bildes scheint die fast senkrechte Kante etwas zu „zerlaufen“ (siehe rechte Vergrößerung in Abbildung 6.16(e)). Durch den Tiefpasseffekt der Extrapolation wird dieser Fehler subjektiv kaum wahrgenommen. Bei Betrachtung des Stereobildpaares (original und virtuell) kann er sogar vollständig vernachlässigt werden. Der Versatz von Kanten, wie er bei der Spiegelung auftreten kann, wird hingegen beim stereoskopischen Sehen verstärkt wahrgenommen.

Die gesamten Berechnungsschritte von RSVS sind in Algorithmus 10 zusammengefasst.

6.2.5. Super-Resolution Stereobildsynthese

In den voran gegangenen Abschnitten wurden die grundlegenden Schritte zur Bestimmung virtueller Stereoansichten mittels RSVS beschrieben. Es wurde gezeigt, dass für die

Ziel: Synthese virtueller Stereoansichten.

- Initialisierung: Vorgabe der Entfernung t_s zwischen dem Weltkoordinatensystem und der ersten Kamera und Wahl der Null-Parallaxenebene Z_C in Metern.
- Normierung der Kamerapositionen mit Gl. 6.15 in Metern.
- **Wiederhole** für alle Bilder der Videosequenz k
  1. Bestimmung der Position der jeweils zugehörigen virtuellen Stereokamera mit Gl. 6.16.
  2. Berechnung der Projektionsmatrix $P^m_{stereo,k}$ der virtuellen Stereokamera mit Gl. 6.17.
  3. Projektion der 3D-Punkte $\mathbf{M}_i$ in die Bildebenen der virtuellen und Originalkameras mit Gl. 6.18 und Gl. 6.19.
  4. Bestimmung der Abstände der Originalkameras $\mathbf{C}^m_k$ zu der virtuellen Kamera $\mathbf{C}^m_{stereo,k}$.
  5. Berechnung der Homographie zwischen der virtuellen Kamera und der dichtesten Originalkamera über Gl. 6.20.
  6. Invertierung der Homographie und Transformation der Farbwerte aus dem Originalbild in die Bildebene der virtuellen Kamera mit Gl. 6.22.
  7. **Wiederhole** solange $\eta < 0.995$ und $\epsilon < 0,6$
     a) Berechnung der Homograpien zwischen virtueller Kamera und der nächst dichtesten Originalkamera über Gl. 6.20.
     b) Invertierung der Homographie und Transformation der Farbwerte aus dem Originalbild in noch nicht gefüllte Bildbereiche der virtuellen Kamera mit Gl. 6.22.

     Ende der Schleife
  8. Extrapolation der nicht aufgefüllten Randbereiche.

Ende der Schleife

Ergebnis: Virtuelle Stereoansichten für jedes Bild der Sequenz.

Algorithmus 10: Realistische Stereobildsynthese (RSVS)

Synthese einer virtuellen Ansicht mehrere umliegende Originalkameraansichten verwendet werden müssen (siehe Abbildung 6.15). Diese Originalansichten besitzen untereinander große Überlappungsbereiche und somit Redundanz, die ausgenutzt werden kann, die Auflösung sowohl der virtuellen Ansichten als auch der Originalansichten zu erhöhen [79, 100].

Super-Resolution durch Ausnutzung räumlicher und zeitlicher Redundanz ist ein weit verbreitetes Forschungsfeld, da damit die Hoffnung verbunden ist, inhärente Beschränkungen aufgenommener Bilder mit geringer Auflösung (engl. *low resolution*, LR) zu verbessern [16, 79, 136]. Das wesentliche Ziel ist es, ein oder mehrere hochauflösende Bilder (engl. *high resolution*, HR) durch Verarbeitung mehrerer LR Bilder, die von unterschiedlichen Kameras oder - in diesem Fall - zu unterschiedlichen Zeitpunkten von einer Kame-

ra aufgenommen wurden, zu rekonstruieren. Der Prozess der Rekonstruktion lässt sich durch Inversion des Beobachtungsmodells [136], welches LR Bilder und HR Bilder in Beziehung setzt, erreichen und beinhaltet in der Regel drei Arbeitsschritte: Registrierung, Interpolation und Restauration.

Ohne Beschränkung der Allgemeingültigkeit wird im Folgenden die Erzeugung eines hochauflösenden virtuellen Stereobildes beschrieben. Die Berechnungsschritte lassen sich ebenso auf die Rekonstruktion hochauflösender Originalansichten anwenden.

Registrierung

Die Berechnung eines virtuellen HR Stereobilds ist abhängig von der gewünschten Auflösung. Je nach Auflösung und Anzahl der Originalansichten in der Videosequenz können nur begrenzt höhere Auflösungen erzielt werden. Beispielhaft soll hier die Auflösung bezüglich der Originalauflösung um den Faktor 1,5 in horizontaler und vertikaler Richtung erhöht werden, d.h. liegt das Eingangsvideo im PAL-Fernsehformat (720×576 Pixel) vor, so ergibt sich für das resultierende Ausgangsvideo eine Auflösung von 1080×864 Pixel.

Der Registrierungsprozess entspricht dabei weitestgehend dem in Abschnitt 6.2.3 beschriebenen Ablauf. Um jedes Pixel des virtuellen HR Stereobildes zu rekonstruieren, müssen korrespondierende Positionen in benachbarten LR Originalansichten ermittelt werden. Dazu werden wiederum die Homographien zwischen der virtuellen HR Stereoansicht und den räumlich benachbarten LR Originalansichten über die korrespondierenden Merkmalspunkte entsprechend Gl. 6.20 bestimmt. Die Position jedes Pixels des HR Bildes in den Originalansichten lässt sich durch Inversion der Homographie mit Gl. 6.21 berechnen. Dabei tritt erneut das Problem auf, dass die ermittelten Positionen in den benachbarten Bildern nicht direkt auf dem Pixelraster liegen, d.h. keine ganzzahligen Werte annehmen, und somit Interpolationsverfahren verwendet werden müssen, um den entsprechenden Farbwert zu erhalten.

Interpolation

Da die Anwendung von Interpolationsverfahren den Effekt der Tiefpassfilterung aufweist - dieser Effekt ist vor allem bei der Vergrößerung eines Bildes durch Hochtastung mit Interpolation benachbarter Pixel desselben Bildes zu beobachten -, ist es wünschenswert, dass korrespondierende Pixel in den Originalansichten möglichst dicht bzw. genau auf dem Pixelraster liegen. Bei Verwendung mehrerer benachbarter Originalansichten ist davon auszugehen, dass sich diese Positionen unterschiedlich nah am Pixelraster befinden. Abbildung 6.17 verdeutlicht diese Annahme.

Die Position eines Pixels im virtuellen HR Stereobild wird über die ermittelten Homographien auf den benachbarten LR Originalansichten abgebildet. In Bild n liegt diese

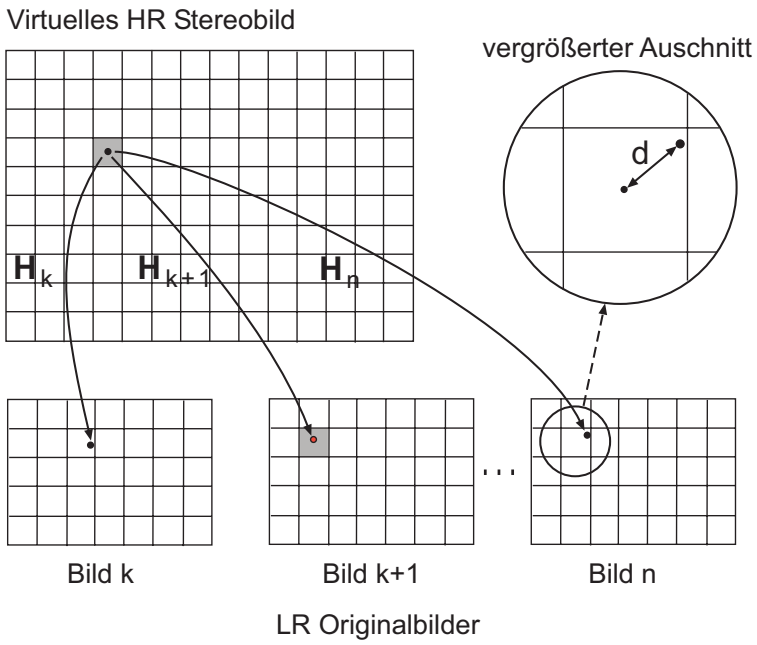


Abbildung 6.17.: Super-Resolution durch räumliche und zeitliche Redundanz

Abbildung zwischen zwei benachbarten Pixeln (siehe Vergrößerung). In Bild $k+1$ hingegen fällt die berechnete korrespondierende Position fast genau auf ein Pixel.

Der Farbwert eines Pixels in der virtuellen HR Stereoansicht lässt sich wieder gemäß Gl. 6.22 unter Verwendung von Interpolationsverfahren durch benachbarte Pixel (siehe Abschnitt 5.2.4) in einer der benachbarten LR Originalansichten ermitteln. Damit die Auswirkungen der Interpolation (Tiefpassfilterung) möglichst gering ausfallen, ist es naheliegend, die Farbinformation aus einem benachbarten Originalbild zu verwenden, bei dem die korrespondierende Position möglichst dicht am Pixelraster liegt (Bild $k+1$ in Abbildung 6.17). Diese Kameraansicht muss nicht zwingend die räumlich dichteste sein, wie es in Abschnitt 6.2.4 gefordert wurde.

Auswahl geeigneter Pixel für die Transformation

Der vorherige Abschnitt gibt zu erkennen, dass ein Pixel, welches am dichtesten am Pixelraster in einer der benachbarten LR Originalansichten liegt, am besten für die Transformation in die virtuelle HR Stereoansicht geeignet ist. Diese Aussage ist allerdings nicht allgemeingültig, da Transformationsfehler bei Zunahme des Kameraabstands größer werden. Zudem können weitere Faktoren, die bei der Aufnahme einer realen Szene mit einer realen Kamera existieren, die Farbwerte einzelner Pixel stark verändern. Darunter fallen z. B. Faktoren wie Beleuchtungsänderungen während der Aufnahme, die in Abhängigkeit von der Zeit und damit mit dem Kameraabstand wahrscheinlicher werden. Zudem

können Lichtreflexionen bzw. Verdeckungen von 3D-Punkten bei unterschiedlichen Kameraperspektiven verstärkt auftreten und somit falsche Farbwerte an einigen Positionen im entsprechenden Originalbild hervorrufen.

Interne Kameraeigenschaften spielen ebenso eine entscheidende Rolle. Z. B. werden Videoaufnahmen meist schon kodiert auf dem Speichermedium der Kamera abgespeichert, wodurch je nach Kompressionsgrad, Bewegung der Kamera und Bildinhalt unterschiedlich stark mit Kodierungsartefakten zu rechnen ist. Zudem kann durch schnelle Bewegung der Kamera (bzw. in der Szene) Bewegungsunschärfe entstehen und somit einzelne Bilder bzw. Teilbereiche der Bilder für eine Transformation ungeeignet machen.

Um diese Faktoren weitestgehend zu umgehen, ist es sinnvoll, Nebenbedingungen bzw. Beschränkungen für die Auswahl und Transformation eines Pixels festzulegen. Eine Beschränkung lässt sich über den Transformationsfehler gemäß Gl. 6.23 realisieren, so dass Kameraansichten mit zu großem Kameraabstand nicht berücksichtigt werden. Als Nebenbedingung für alle in Frage kommenden benachbarten Kameraansichten gilt die Eigenschaft, dass alle zugeordneten Pixelpositionen in den Originalansichten einen ähnlichen Farbwert aufweisen müssen. Weichen einzelne Farbwerte zu stark vom Medianwert aller Farbwerte an den entsprechenden Positionen ab, so sind diese Pixelzuordnungen statistisch unglaubwürdig, auch wenn sie dicht am Pixelraster liegen sollten.

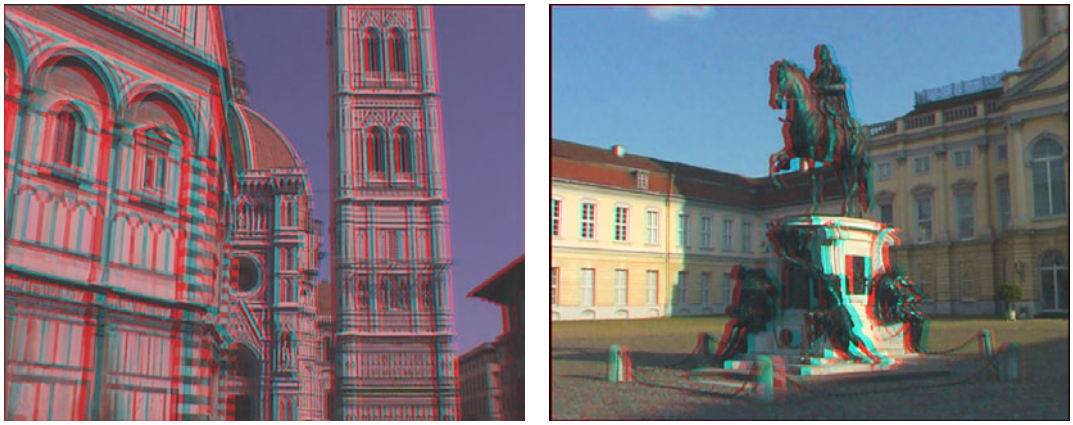
Für die berechneten Pixelpositionen wird demnach jeweils mittels Interpolation der entsprechende Farbwert ermittelt. Anschließend lässt sich mit

$$I_{med}\left(\mathcal{T}(x,y)\right) = \underset{\forall k}{\mathbf{median}}\, I_k\left(\mathcal{T}_k(x,y)\right) \tag{6.24}$$

der Medianwert bestimmen, wobei I_k der Farbwert des Pixels im Originalbild k ist. Farbwerte, die eine zu hohe Abweichung vom Medianwert besitzen, werden verworfen, d.h. nicht für die Transformation in die virtuelle HR Stereoansicht berücksichtigt. Für alle übrigen Farbwerte wird derjenige entsprechend in der virtuellen Stereoansicht eingesetzt, dessen Pixelposition am dichtesten am Pixelraster liegt, also möglichst geringe Auswirkungen durch die Interpolation erfahren hat.

6.2.6. Ergebnisse

In diesem Abschnitt werden beispielhaft die Ergebnisse einiger mittels RSVS generierter Stereo- und Multiview-Ansichten sowie Ansichten, deren Auflösung mittels Super-Resolution erhöht wurden, dargestellt. Die Evaluierung eines umfangreichen Datensatzes konvertierter Sequenzen wird in Kapitel 7 durchgeführt.



(a) (b)

Abbildung 6.18.: Anaglyph Stereobilder (Handkamera) der Sequenzen (a) *Dom*, (b) *Statue*

Tabelle 6.1.: Entfernungsangaben und durchschnittliche Anzahl verwendeter Originalansichten, um eine virtuelle Stereoansicht zu erzeugen

Sequenz	Auflösung in Pixeln	t_s in m	Z_C in m	durchschnittliche Anzahl verwendeter Ansichten
Medusa	448 x 358	3	6	14,04
Dom	720 x 576	15	8	61,24
Fassade	720 x 576	12	10	37,11
Statue	720 x 576	8	10	53,70
Kirche	576 x 720	12	16	8,14

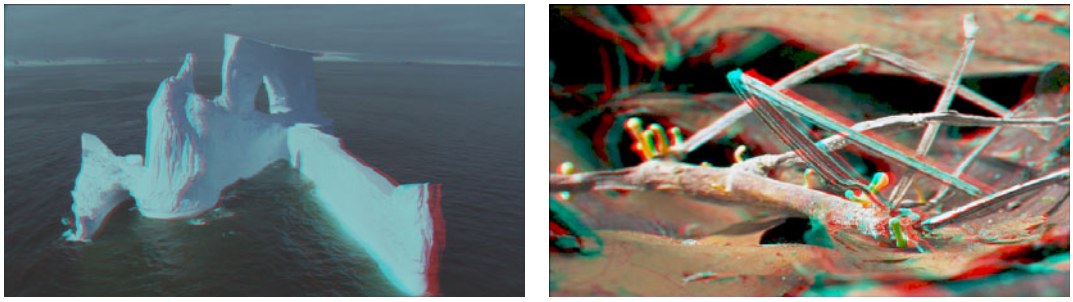
Standard RSVS (Stereo)

Abbildung 6.18 zeigt zwei mittels RSVS erzeugte Stereobilder (Anaglyphenbilder[4]) aus den mit einer Handkamera aufgenommenen Videosequenzen *Dom* und *Statue*. In Tabelle 6.1 sind die relevanten Einstellungen wie Bildauflösung, Entfernung der Szene (t_s) und Entfernung der Null-Parallaxen-Ebene (Z_C) sowie die durchschnittlich für die Synthese einer virtuellen Stereoansicht benötigte Anzahl an Originalansichten angegeben.

Bei Handkameras ist die Kameraführung in der Regel sehr schwankend, wodurch die Orientierung und somit das Sichtfeld der Kamera stark variieren kann. Wird eine parallele virtuelle Stereokamera definiert, kann diese sowohl horizontal als auch vertikal rotiert gegenüber der dichtesten Originalkamera sein, die für die Synthese verwendet wird, und somit unterschiedliche Bereiche der Szene aufnehmen. Um die nötigen Bildinformationen der virtuellen Kamera möglichst vollständig zu erhalten, müssen gerade bei handgeführten Aufnahmen verstärkt zusätzliche Originalansichten verwendet werden.

Aus der BBC-Dokumentation *Planet Erde* wurden über 60 Videosequenzen (insgesamt ca. 13 Minuten) mittels RSVS in Stereosequenzen konvertiert. Abbildung 6.19 zeigt bei-

[4] Anaglyphenbilder sind überlagerte stereoskopische Bilder, die in Komplementärfarben (hier: rot-cyan) dargestellt werden



(a) (b)

Abbildung 6.19.: Anaglyph Stereobilder (BBC) der Sequenzen (a) *Eisberg*, (b) *Insekt* (Quelle: BBC-Dokumentation *Planet Erde*)

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

Abbildung 6.20.: Multiview Bildpaare (Handkamera) der Sequenz *Statue*: (a) $t_x = -128mm$, (b) $t_x = -64mm$, (c) Originalansicht, (d) $t_x = 64mm$ und (e) $t_x = 128mm$

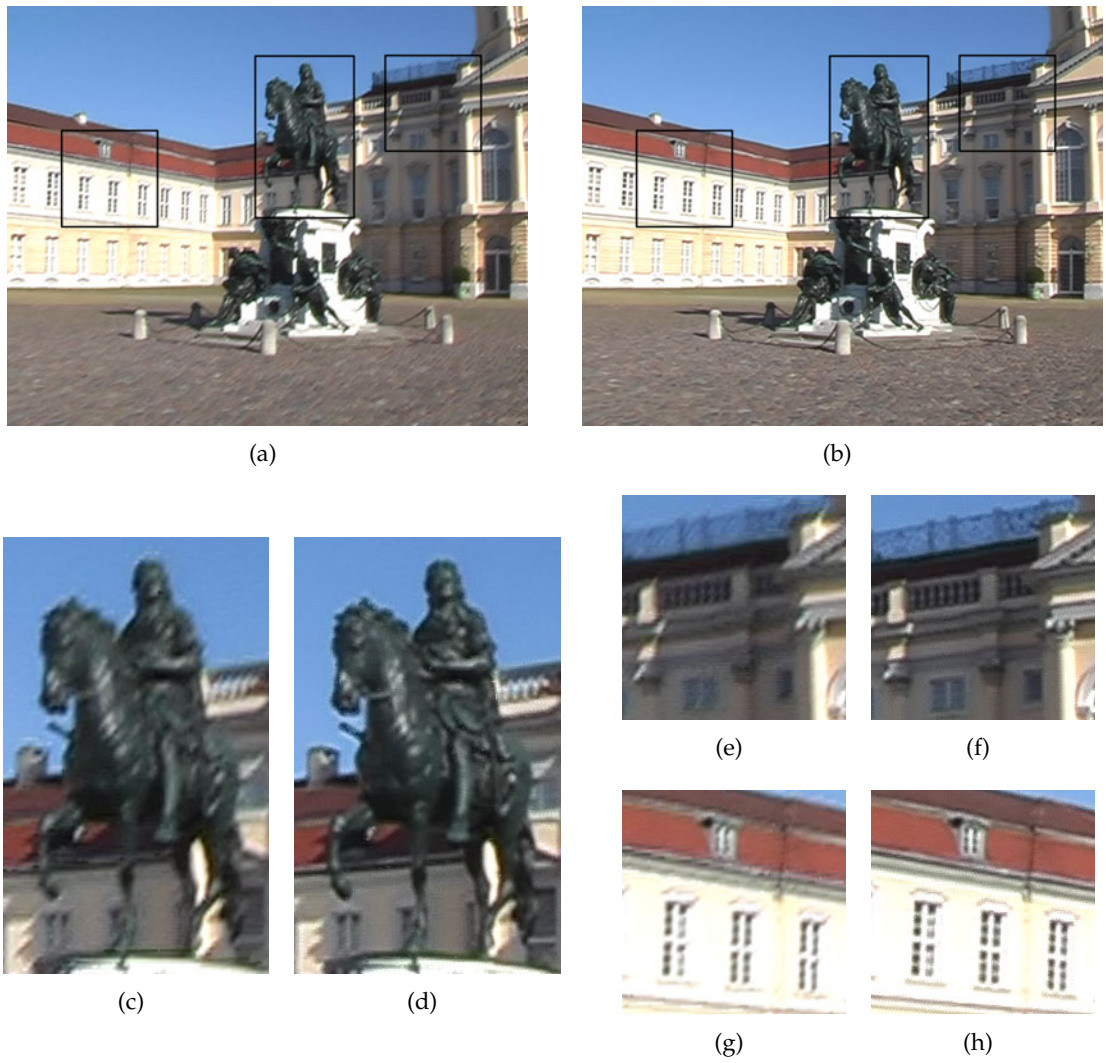
(a) (b) (c) (d) (e)

Abbildung 6.21.: Multiview Bildpaare (BBC) der Sequenz *Gebirge*: (a) $t_x = -128mm$, (b) $t_x = -64mm$, (c) Originalansicht, (d) $t_x = 64mm$ und (e) $t_x = 128mm$ (Quelle: BBC-Dokumentation *Planet Erde*)

spielhaft zwei generierte Stereobilder (Anaglyphenbilder) aus diesem Datensatz. Für die Sequenz *Eisberg* wurden durchschnittlich 7,69 Originalansichten pro virtuelles Stereobild benötigt. Bei der Sequenz *Insekt* waren es 9,23 Originalansichten pro Stereobild.

Standard RSVS (Multiview)

RSVS ist nicht darauf limitiert, nur eine einzelne Stereoansicht für jedes Bild der Videosequenz zu erzeugen. Durch Definition eines virtuellen Multiview-Kamerasetups, wie in Abbildung 6.12 dargestellt, lassen sich mehrere virtuelle Ansichten, die für autostereoskopische Displays benötigt werden, generieren. Die Abbildungen 6.20 (Handkamerasequenz *Statue*) und 6.21 (BBC-Sequenz *Gebirge*) zeigen jeweils vier mittels RSVS erzeugte virtuelle Ansichten und die Originalansicht in der Mitte.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Abbildung 6.22.: Vergleich zwischen Standard RSVS und RSVS mit Super-Resolution für die Sequenz *Statue*: (a) Hochtastung mit Lanczos-Filterung (1080 × 864 Pixel), (b) Super-Resolution (1080 × 864 Pixel), (c), (e), (g) vergrößerter Ausschnitt aus (a), (d), (f), (h) vergrößerter Ausschnitt aus (b).

RSVS mit Super-Resolution

Zwei Beispielbilder aus den Handkamerasequenzen *Statue* und *Dom* (jeweils *interlaced* PAL Aufnahmen) sollen die Leistungsfähigkeit der Erweiterung von RSVS mit Super-Resolution verdeutlichen. In Abbildung 6.22 werden die Ergebnisse der Synthese einer stereoskopischen Ansicht mittels Standard RSVS und RSVS mit Super-Resolution für die Sequenz *Statue* gegenübergestellt. Die Auflösung für beide Verfahren wurde um den Faktor 1,5 auf 1080 × 864 Pixel erhöht. Bei Standard RSVS ist diesbezüglich eine Hochtastung mit Lanczos-Filterung durchgeführt worden.

Abbildung 6.22(g) zeigt einige typische Artefakte von *interlaced* PAL Video und anschließender Hochtastung: Entlang der Kanten sind Sägezahnmuster zu erkennen, die

durch *Deinterlacing* verursacht wurden. Bei der Super-Resolution (Abbildung 6.22(h)) ist dies nicht zu beobachten. Zudem liefert Super-Resolution zwei zusätzliche Vorteile gegenüber Hochtastung des virtuellen Stereobildes: Schatten an Objektkanten (Geisterbilder), verursacht durch die Videokompression (siehe Reiterstatue und Schornstein in Abbildung 6.22(c)), und Bewegungsunschärfe durch unruhige Kameraführung (siehe Abbildung 6.22(c) und 6.22(e)) werden mit Super-Resolution stark reduziert.

In Abbildung 6.23 sind die Ergebnisse von Standard RSVS und RSVS mit Super-Resolution für die Sequenz *Dom* dargestellt. Das Sägezahnmuster an den Objektkanten (siehe Abbildung 6.23(c) und 6.23(g)) konnte durch Super-Resolution stark verringert werden. Allerdings ist bei der Super-Resolution geringfügig das Auftreten des *Zipper*-Effekts (*Reißverschluss*-Effekt verursacht durch geringe Registrierungsfehler) zu beobachten (siehe Abbildung 6.23(d) an den Kanten des Fensterrahmens).

Ein weiterer Vorteil der Super-Resolution ist die Reduzierung von Spiegelungsfehlern durch Unterabtastung (engl. *Aliasing*) in hochfrequenten Bildbereichen (siehe oberer Fensterbogen in Abbildung 6.22(g) und 6.22(h)).

6.3. Kapitelzusammenfassung und Diskussion

In diesem Kapitel wurden zwei Verfahren (DIBR und RSVS) zur Synthese virtueller Ansichten vorgestellt. Da DIBR dichte Tiefenkarten benötigt, wurde einführend ein Überblick über verschiedene Verfahren zur Bestimmung von Tiefenkarten gegeben. Die grundlegende Funktionsweise mit den Problemen von DIBR wurde anschließend beschrieben und zusätzliche Modifikationen wie die Vorverarbeitung der Tiefenkarten und der Einsatz des Shift-Sensor-Ansatzes näher erläutert.

Der zweite Teil dieses Kapitels widmete sich einem neuen Ansatz zur Synthese virtueller Ansichten mittels RSVS. Dieser benötigt zwar keine dichten Tiefenkarten, unterliegt aber Einschränkungen bezüglich Kameraführung bei der Aufnahme. RSVS setzt direkt auf die Ergebnisse von SfM auf und berechnet nach einer Initialisierung Stereo- oder Multiview-Ansichten vollständig automatisch. Ferner wurde eine Erweiterungsmöglichkeit von RSVS vorgestellt, die mithilfe von Super-Resolution die Auflösung sowohl der virtuellen als auch der Originalansichten erhöhen kann.

Abschließend wurden die Ergebnisse der Synthese einiger stereoskopischer und Multiview-Ansichten mittels RSVS dargestellt. Als Datensatz lagen Videosequenzen, die mit einer Handkamera aufgenommen wurden, und Sequenzen aus der BBC-Dokumentation *Planet Erde* vor. Für zwei Handkamerasequenzen wurde die Erweiterung durch Super-Resolution mit Standard RSVS verglichen und gezeigt, dass Artefakte wie Sägezahnmuster an Kanten, Kodierungsartefakte, Bewegungsunschärfe und Aliasing zum Teil stark reduziert werden konnten.

Abbildung 6.23.: Vergleich zwischen Standard RSVS und RSVS mit Super-Resolution für die Sequenz *Dom*: (a) Hochtastung mit Lanczos-Filterung (1080 × 864 Pixel), (b) Super-Resolution (1080 × 864 Pixel), (c), (e), (g) vergrößerter Ausschnitt aus (a), (d), (f), (h) vergrößerter Ausschnitt aus (b).

Der vorgestellte RSVS Ansatz unterliegt allerdings einigen Einschränkungen. Die wesentlichste Einschränkung ist, dass die aufgenommene Szene statisch sein muss, d.h. zusätzlich bewegte Objekte innerhalb der Szene würden die stereoskopische Tiefenwahrnehmung negativ beeinflussen. Zudem unterliegt die Kamerabewegung gewissen Beschränkungen. Bei einer vollständig vorwärts- oder rückwärts gerichteten Kamerabewegung würde die virtuelle Stereokamera zu weit vom Kamerapfad entfernt liegen, d.h. die dichteste Originalkamera, die vorrangig für die Synthese verwendet wird, würde der zugeordneten Originalstereoansicht entsprechen. Da dadurch keine zusätzlichen Informationen

aus einer anderen Perspektive vorhanden sind, entstünde lediglich eine perspektivische Verzerrung der Originalansicht in der virtuellen Kamera.

Ferner unterliegt die Einstellung der Parallaxe (durch variieren von t_s und Z_C) gewissen Einschränkungen. Durch Erhöhung der Parallaxe und damit Verstärkung des 3D-Effekts wird indirekt die Basislinienlänge zwischen Originalkamera und paralleler virtueller Stereokamera vergrößert. Je größer dieser Abstand aber wird, umso größer ist auch die Wahrscheinlichkeit, dass sich die virtuelle Kamera zu weit vom Kamerapfad entfernt. Dieser Fall tritt vor allem bei der Erzeugung weiter entfernter Multiview-Ansichten verstärkt auf (siehe Vergrößerung unten links in Abbildung 6.12). Eine 2D-Transformation wäre hier aufgrund zu starker perspektivischer Verzerrungen ggf. nicht mehr in einem tollerierbaren Rahmen.

Der entwickelte RSVS Ansatz ist neuartig und enthält Vor- und Nachteile gegenüber Verfahren des aktuellen Forschungsstands. Nachteilig sind die aufgeführten Beschränkungen von RSVS, die allerdings auch bei dem Verfahren in [152] vorhanden sind. Zu den wesentlichen Vorteilen zählen die fotokonsistenten Eigenschaften des Ansatzes, da ausschließlich Originalansichten über perspektivische Transformationen verwendet werden. Durch die genaue Definition der virtuellen Kameras werden die Transformationen direkt auf die Bildebene und nicht wie in [152] auf eine dominante Ebene in der Szene angewendet. Zudem wird damit ein zeitlich konsistenter Tiefeneindruck erzeugt.

Aufgedeckte Bereiche innerhalb der virtuellen Ansicht, wie sie z. B. bei DIBR auftreten, sind bei RSVS nicht bzw. nur im Randbereich vorhanden. DIBR benötigt im Gegensatz zu RSVS dichte Tiefenkarten. Die Berechnung dieser Tiefeninformation ist sehr aufwendig und fehleranfällig, wie in den 2D/3D-Konvertierungsansätzen in Abschnitt 6.1.4 gezeigt wurde, wodurch zusätzliche Fehler bei der Synthese auftreten.

Ein weiterer Vorteil von RSVS ist, dass durch die Verwendung mehrerer zeitlich und räumlich versetzter Originalansichten zur Synthese einer virtuellen Ansicht die Auflösung erhöht werden kann. Die Konvertierung monokularer Videosequenzen in stereoskopische Sequenzen mit erhöhter Auflösung durch den Einsatz von Super-Resolution wurde in der Literatur bislang noch nicht behandelt.

Abgesehen von den beschriebenen Einschränkungen ist RSVS ein sehr leistungsfähiges Verfahren, welches als Unterstützung für benutzergesteuerte 2D-3D-Konvertierung oder 3D-Produktionssysteme verwendet werden kann. Qualitativ hochwertige Konvertierungen und Postproduktion wird derzeit noch mit manuellen oder semiautomatischen Softwareprodukten durchgeführt (wie z. B. in [63]). RSVS als automatisches Konvertierungsverfahren kann für einzelne Sequenzen (Teile von Filmen) den manuellen Aufwand stark reduzieren.

7. Evaluierung

Die in Kapitel 6 vorgestellten Syntheseverfahren, DIBR [44] und RSVS, sollen nun hinsichtlich ihrer Qualität untersucht werden. Objetive Qualitätsmessungen, wie sie z. B. bei der Videokodierung in Form von PSNR-Berechnung verwendet werden, sind bei der Synthese neuer Ansichten ungeeignet, da sie keinerlei Aussagekraft besitzen. Zwar lassen sich Referenzaufnahmen mit stereoskopischen Aufnahmeeinrichtungen erzeugen, allerdings kann schon bei den Analyseverfahren die exakte Berechnung einer virtuellen Kamera an der Stelle der Referenzkamera nicht gewährleistet werden. Ein Versatz des Bildes um ein oder mehrere Pixel würde den PSNR drastisch reduzieren, obwohl gleichzeitig die Bildqualität und die stereoskopische Tiefenwahrnehmung zunehmen kann. Subjektive Qualitätstests, die ebenfalls gerade bei Videokodierungsverfahren Anwendung finden [139], besitzen dagegen eine hohe Aussagekraft über die tatsächliche Qualität der verwendeten Syntheseverfahren. Für die Videokodierung existieren bereits hinreichend Standards - wie ITU-R BT.500-10 [81] - für die subjektive Beurteilung der Qualität und werden in Teilen ebenso für die Stereoskopie übernommen [82]. Innerhalb des Europäischen Forschungsprojekts „3DTV“[1] wurde festgestellt, dass die Standardisierung für 3D-Anwendungen noch nicht ausreichend ist und weiter vorangetrieben werden muss, da z. B. zusätzliche psychovisuelle Beeinträchtigungen in stereoskopischen Videosequenzen auftreten können [125].

7.1. Überblick über die Evaluierungsmethoden

In den folgenden beiden Abschnitten werden zwei subjektive Tests durchgeführt, um die Qualität von RSVS zu beurteilen und einen Vergleich zwischen DIBR und RSVS vornehmen zu können. Die Testbedingungen sind standardkonform mit den ITU-Empfehlungen in [81, 82]. Da keinerlei 2D-Referenzdatensätze mit Referenztiefenkarten bzw. stereoskopische Sequenzen, die als Referenz verwendet werden können, vorliegen[2], wird in dem ersten Experiment lediglich die Qualität von RSVS mittels der *Single-Stimulus* (SS) Methode [139] untersucht (siehe Abschnitt 7.2.1). Zum Vergleich der Qualität dienen die 2D-Originalsequenzen und stereoskopische Sequenzen, die durch einen Zeitversatz der Originalsequenzen erzeugt wurden [151]. Für das zweite Experiment wurde mit *Autodesk*

[1] Network of Excellence 3DTV (FP6-PLT-511568-3DTV)

[2] Es existieren zwar eine Reihe von Testdatensätzen, allerdings wurden diese mit einer statischen Kamera und Zcam bzw. mit einem statischen Multiview-Kamerasetup aufgenommen und sind damit nicht für RSVS geeignet.

3ds Max ein 3D-Modell erstellt und mehrere synthetische Sequenzen inklusive pixelgenauer Tiefenkarten mit Hilfe einer Stereokamera gerendert. Da damit sowohl stereoskopische Referenzdatensätze als auch exakte Tiefenkarten zur Verfügung stehen, kann ein Vergleich von DIBR und RSVS mittels der *Double-Stimulus-Continous-Quality-Scale* (DSCQS) Methode [82,139] vorgenommen werden (siehe Abschnitt 7.3.1). Für beide Experimente wurden jeweils 15 Testpersonen ausgewählt, die die Qualtität der gezeigten Sequenzen mit Hilfe von Bewertungsskalen beurteilen sollten. Direkt im Anschluss an die Tests wurden den Testpersonen zusätzlich zwei Fragen gestellt: 1. „Ist Ihnen bei dem Test irgendetwas Besonderes aufgefallen?“ und 2. „Welche Störungen bzw. Fehler sind Ihnen aufgefallen?“. Eine detaillierte Beschreibung der Versuchsdurchführungen und der verwendeten Evaluierungsmethoden ist in den folgenden Abschnitten zu finden.

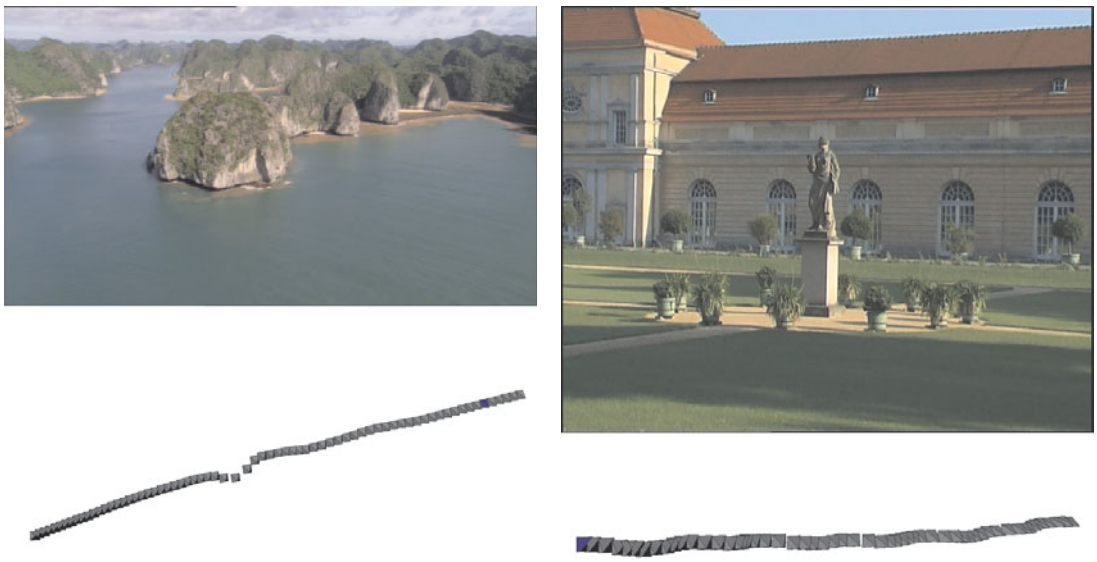
7.2. Subjektiver Qualitätstest I (SS-Methode)

In dem ersten Experiment wurden 15 fachfremde Testpersonen aufgefordert, den Qualitätseindruck von 25 Testsequenzen (18 Sequenzen aus der BBC-Dokumentation *Planet Erde* mit gleichmäßiger und überwiegend horizontaler Kameraführung und sieben Handkamerasequenzen mit schwankender aber ebenfalls überwiegend horizontaler Kameraführung) zu bewerten. Die Handkamerasequenzen wurden mit einer Canon XLs 1 (progressive Aufnahme, DV kodiert) aufgenommen. In Abbildung 7.1 sind zwei Beispielbilder aus den Testdatensätzen inklusive Kamerapfad dargestellt. Die blaue Kamera gibt jeweils die Position an, von der aus das darüber befindliche Bild aufgenommen wurde. Die Testdatensätze unterscheiden sich vor allem in der Kameraführung. Bei den Handkamerasequenzen ist die Kameraführung sehr schwankend und unstetig. Für alle Sequenzen gilt, dass die aufgenommenen Szenen überwiegend statische Landschaftsaufnahmen sind.

Jede der Testsequenzen hatte eine Länge von ca. 10 Sekunden und wurde mit variierender Reihenfolge in drei verschiedenen Modi dargestellt:

1. 2D (die Originalansicht wurde sowohl für das linke als auch für das rechte Auge verwendet)
2. 3D mittels RSVS
3. 3D durch Zeitversatz (d.h. die Stereosequenz wurde durch einen Versatz der Originalsequenz, der zwischen 4 und 10 Bildern lag, erzeugt)

Vor dem jeweiligen Test wurde dem Probanden ein Informationsblatt über die Testdurchführung (siehe Anhang A.25) ausgehändigt.



(a) (b)

Abbildung 7.1.: Beispielbilder der Testsequenzen für Qualitätstest I: (a) Sequenz *Inselgruppe* (Quelle: BBC-Dokumentation *Planet Erde*) mit Kamerapfad und (b) Sequenz *Schlossgarten* mit Kamerapfad.

7.2.1. Single-Stimulus Methode

Zur Durchführung des Tests wurde die Single-Stimulus Methode (SS-Methode) mit diskreter Bewertungsskala verwendet, d.h. nach jeder gezeigten Testsequenz (diese wurde nur einmal dargestellt) standen dem Probanden 10 Sekunden Zeit zur Verfügung, ein Kreuz an der entsprechenden Stelle in der Bewertungsskala zu platzieren. Innerhalb dieser 10-sekündigen Bewertungsphase wurde ein Bild mit mittlerem Grauwert ausgestrahlt. Die Bewertungsskala für dieses Experiment ist in Abbildung 7.2(a) dargestellt.

Der Test bestand aus zwei Phasen: einer kurzen Trainingsphase und einer ca. 27-minütigen Evaluierungsphase. Sowohl bei der Trainings- als auch bei der Evaluierungsphase musste der Proband eine Bewertung vornehmen. Bei Unregelmäßigkeiten bzw. auftretenden Unklarheiten während der Trainingsphase war diese zu wiederholen.

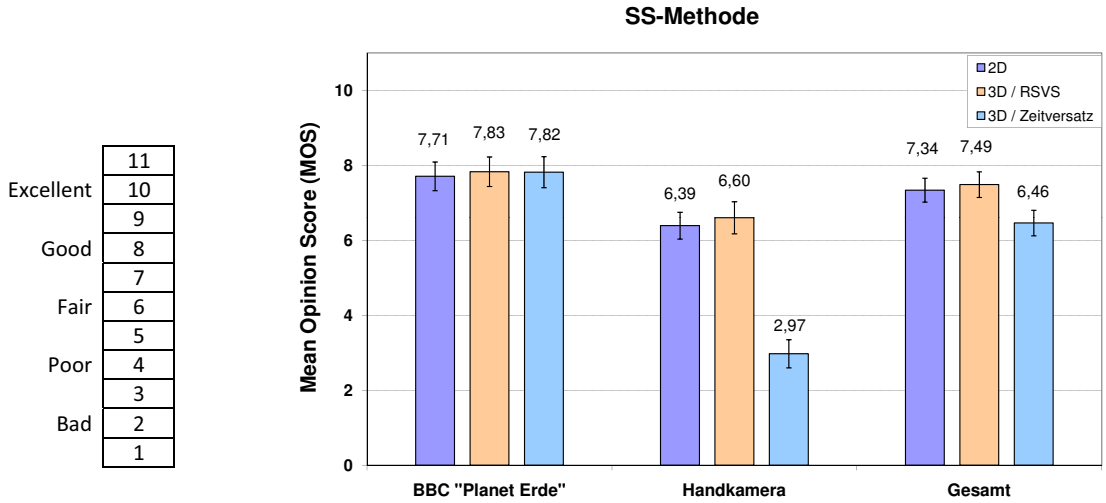
Als Bewertungskriterium sollte lediglich der qualitative Gesamteindruck der gezeigten Testsequenzen beurteilt werden. Der Hinweis auf mögliche Artefakte wäre nur auf die 3D-Modi anwendbar gewesen und hätte damit bereits im Voraus diese suggestiv beeinflusst. Zudem wären beide 3D-Modi damit im Vorfeld degradiert worden, da im 2D-Modus diese möglichen Artefakte nicht auftreten.

7.2.2. Auswertung

Der durchschnittlich erzielte subjektive Qualitätseindruck (engl. *mean opinion score*, MOS) und die Standardabweichung für die beiden Testdatensätze sind in Tabelle 7.1 dargestellt.

Tabelle 7.1.: Subjektiver Qualitätstest I (SS-Methode): Durchschnittlich erzielter subjektiver Qualitätseindruck (MOS) und Standardabweichung.

	2D		3D / RSVS		3D / Zeitversatz	
Datensatz	MOS	±Std.Abw.	MOS	±Std.Abw.	MOS	±Std.Abw.
BBC	7,707	1,663	**7,830**	1,703	7,819	1,792
Handkamera	6,390	1,550	**6,600**	1,858	2,971	1,632
Gesamt	7,339	1,631	**7,485**	1,746	6,461	1,747



(a) (b)

Abbildung 7.2.: Subjektiver Qualitätstest I (SS-Methode): (a) Bewertungsskala, (b) Durchschnittlich erzielter subjektiver Qualitätseindruck (MOS) mit 95% Konfidenzinterval.

Zur besseren Anschauung dient das Diagramm in Abbildung 7.2(b), welches ebenfalls den MOS mit 95% Konfidenzintervall für alle Testsequenzen zeigt.

Während des Tests sind zwei Unregelmäßigkeiten aufgetreten. Zum einen sollte der Test mit einer Shutter-Brille durchgeführt werden. Diese hatte Synchronisierungsprobleme und konnte somit nicht verwendet werden. Das zur Verfügung stehende autostereoskopische Display mit Kopfverfolgung (engl. *head tracking*) war für die Tests ebenso ungeeignet, da durch geringfügige Kopfbewegungen ein Pseudo-3D-Effekt[3] entstehen konnte. Zur Trennung der Stereosequenzen mussten die Probanden letztendlich eine Rot-Cyan-Brille aufsetzen. Bei der zweiten Unregelmäßigkeit hatten zwei der Probanden keine Bewertung während der Trainingsphase vorgenommen. Die Trainingsphase wurde demzufolge jeweils wiederholt.

[3] Vertauschen der linken und rechten Ansicht

Bezüglich der Gesamtbewertung der beiden Datensätze liefert RSVS mit einem MOS von 7,485 den besten Qualitätseindruck. Die Aussagekraft ist allerdings eher gering, da mit den Handkamera- und den BBC-Sequenzen zwei sehr verschiedene Kameraführungen vorlagen. Betrachtet man die Ergebnisse der beiden Datensätze getrennt voneinander, so liefert RSVS zwar jeweils mit einem MOS von 7,83 (BBC) bzw. 6,60 (Handkamera) die besten Ergebnisse, diese allerdings mit geringem Abstand zu mindestens einer der beiden anderen Darstellungsmethoden.

Aufgrund der gleichmäßigen horizontalen und äußerst linearen Kameraführung bei den BBC-Sequenzen liefert ein einfacher Zeitversatz der Sequenzen fast identische Ergebnisse wie RSVS. Da durch einen Zeitversatz ausschließlich Originalansichten dargestellt werden, entstehen hier zumindest keine Einbußen in der Bildqualität. Bei RSVS wird eine virtuelle Ansicht mittels Homographien und Interpolationsverfahren generiert, wodurch leichte Transformationsfehler entstehen können bzw. leichte Tiefpasseffekte auftreten. Vier der Probanden haben z. B. auf die erste Frage („Ist Ihnen bei dem Test irgendetwas Besonderes aufgefallen?“) geantwortet, dass im 3D-Modus teilweise Objekte unscharf wirkten. Die geringfügig schlechtere Bewertung bei „3D / Zeitversatz“ entsteht demnach durch Verschlechterung des Tiefeneindrucks. Keinem der Probanden ist z. B. aufgefallen, dass in ein paar wenigen BBC-Sequenzen beim Zeitversatz eine geringfügige vertikale Parallaxe vorhanden war. Vertikale Parallaxe verursacht allerdings bei längerer Betrachtung eine erhöhte Augenbelastung [75].

Die Darstellungsmethode „2D“, die für beide Datensätze lediglich als Vergleich zu den 3D-Darstellungsmethoden herangezogen worden ist, wurde bei den BBC-Sequenzen mit einem MOS von 7,707 ebenfalls hoch bewertet. Da bei dem Test nicht direkt der 3D-Tiefeneindruck bewertet werden sollte, sondern nur allgemein die Qualität der Sequenzen, war mit diesem hohen Wert zu rechnen. Im 2D-Modus werden sowohl dem linken als auch dem rechten Auge die gleichen Originalsequenzen gezeigt. Störende 3D-Artefakte wie vor allem Nebensprechen (engl. *cross talk*) [125], welches aufgrund der schlechten Trennung der beiden Ansichten durch die Filtereigenschaften der 3D-Brille hervorgerufen wird, sind im 2D-Modus von Natur aus nicht sichtbar. 11 der 15 Probanden antworteten auf die zweite Frage („Welche Störungen bzw. Fehler sind Ihnen aufgefallen?“) im Anschluss an den Test, dass sie „Geisterbilder“ (Nebensprechen) bemerkt haben und dadurch die Qualität geringer bewerteten. Da Nebensprechen lediglich durch die Darstellungsmethode verursacht wird und nicht auf die Synthese zurückzuführen ist, kann demnach generell von einem positiven Offset für 2D ausgegangen werden, d.h. wäre die Kanaltrennung zwischen linker und rechter Ansicht perfekt gewesen, dann ist davon auszugehen, dass der Abstand zwischen der 2D- und 3D-Darstellungsmethode größer ausgefallen wäre.

Bei den Handkamerasequenzen wurde „3D / Zeitversatz“ von allen Probanden als schlecht beurteilt. Da die Kameraführung sehr unstetig war, änderte sich schlagartig sowohl die horizontale als auch die vertikale Parallaxe. Ein Vergleich mit „3D / RSVS“

(a) (b)

Abbildung 7.3.: Beispielbilder für unterschiedliche Tiefenverteilung: (a) Sequenz *Berghang* und (b) Sequenz *Wald* (Quelle: BBC-Dokumentation *Planet Erde*).

ist deshalb nur begrenzt möglich, d.h. der Test soll hier lediglich die zusätzlichen Einsatzmöglichkeiten und die Flexibilität von RSVS im Hinblick auf die Konvertierung von Handkameraaufnahmen demonstrieren.

Der 2D-Modus wurde wie bei den BBC-Sequenzen leicht geringer bewertet als der 3D-Modus mittels RSVS. Hier ist, wie bereits erwähnt, von einem positiven Offset für 2D auszugehen, da Nebensprechen nicht auftritt. Insgesamt gaben fünf Probanden an, dass sie die 3D-Darstellung bevorzugen, da diese durch den Tiefeneffekt einen zusätzlichen plastischen Eindruck der Szene vermittelt.

Weitere Störungen bzw. Fehler, die einigen der Probanden während des Tests aufgefallen sind, waren Kodierartefakte (5 Probanden), teilweise zu starker 3D-Effekt (4 Probanden) und der sogenannte Pappkarton-Effekt (engl. *cardboard effect*) [125] (3 Probanden), bei dem 3D-Objekte flach erscheinen. Der Pappkarton-Effekt, der im eigentlichen Sinne nicht vorhanden war, wurde vermutlich bei denjenigen Sequenzen beobachtet, die einen sehr nahen Vordergrund und einen weit entfernten Hintergrund besaßen, ohne dass zwischen diesen extremen Bereichen Objekte vorhanden waren. Abbildung 7.3 zeigt zwei Beispielbilder der BBC-Sequenzen *Berghang* und *Wald* mit unterschiedlichen Verteilungen der Tiefen. Bei der Sequenz *Berghang* könnte die 3D-Darstellung solch einen Effekt hervorrufen.

7.3. Subjektiver Qualitätstest II (DSCQS-Methode)

In dem zweiten Experiment wurden wiederum 15 fachfremde Testpersonen aufgefordert, den Qualitätseindruck von drei synthetischen Testsequenzen zu beurteilen. Da sowohl stereoskopische Referenzsequenzen als auch exakte Tiefenkarten vorlagen, wurde der Test mit der DSCQS-Methode durchgeführt. In Abbildung 7.4 sind Beispielbilder aus den drei Testsequenzen inklusive Kamerapfad dargestellt. Die blaue Kamera gibt jeweils die Position an, von der aus das darüber befindliche Bild gerendert wurde. Die Testsequenzen

(a) (b) (c)

Abbildung 7.4.: Beispielbilder der Testsequenzen für Qualitätstest II: (a) Sequenz *TUBroom 1*, (b) Sequenz *TUBroom 2* und (c) Sequenz *TUBroom 3* jeweils mit Kamerapfad.

unterscheiden sich vor allem wieder in der Kameraführung, die unterschiedliche Auswirkungen sowohl auf die Konvertierungsmethoden als auch zum Teil auf die Referenzsequenzen hat.

Jede der Testsequenzen hatte eine Länge von ca. 9 Sekunden und wurde mit variierender Reihenfolge in fünf verschiedenen Modi dargestellt:

- Zeitversatz
- Standard RSVS
- RSVS LR
- Standard DIBR
- DIBR LR

„LR“ bedeutet, dass sowohl die linke als auch die rechte Ansicht virtuell erstellt wurden. Der Kameraabstand zwischen den virtuellen Ansichten ist identisch mit dem jeweiligen Standardverfahren. Die Originalkamera liegt in der Mitte zwischen den virtuellen Ansichten, so dass jeweils nur die halbe Parallaxe nach links und nach rechts erzeugt wurde.

Der Vorteil für DIBR LR liegt in der Verringerung der aufgedeckten Bereiche. Allerdings sind diese in beiden Ansichten vorhanden. Zudem entsteht in beiden Sequenzen ein Tiefpasseffekt durch Interpolation. Bei RSVS LR liegen die virtuellen Kameras bei ggf. leicht vorwärtsgerichteter Kameraführung dichter am Kamerapfad, wodurch Transformationsfehler reduziert werden können. Allerdings entsteht hier ebenfalls ein Tiefpasseffekt in beiden Ansichten. Bei beiden DIBR-Verfahren wurde eine Gaussfilterung der Tiefenkarten durchgeführt.

Vor dem jeweiligen Test wurde dem Probanden ein Informationsblatt über die Testdurchführung (siehe Anhang A.26) ausgehändigt.

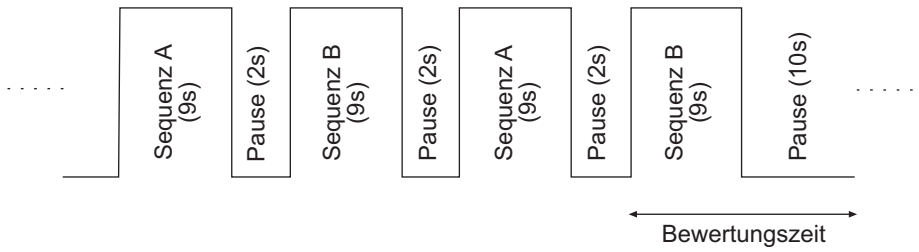


Abbildung 7.5.: Zeitlicher Ablauf der DSCQS-Methode

7.3.1. Double-Stimulus-Continous-Quality-Scale Methode

Zur Durchführung des Tests wurde die Double-Stimulus-Continous-Quality-Scale Methode (DSCQS-Methode) mit kontinuierlicher Bewertungsskala (siehe Abbildung 7.6(a)) verwendet. Sowohl die Testsequenz als auch die Referenzsequenz wurden mit einer kurzen Unterbrechung von 2 Sekunden, in der ein Bild mit mittlerem Grauwert gezeigt wurde, jeweils zwei Mal dem Probanden präsentiert. Dabei ist dem Probanden im Vorfeld nicht bekannt gewesen, welches die Referenz- und welches die Testsequenz war. Für die Bewertung beider Sequenzen, bei der der Proband jeweils einen horizontalen Strich an der entsprechenden Stelle auf beiden Bewertungsskalen (A und B für Test- bzw. Referenzsequenz) platzieren musste, standen die Zeit der letzten Wiederholung der Test- oder Referenzsequenz und zusätzlich eine 10-sekündige Pause, in der wiederum ein Bild mit mittlerem Grauwert gezeigt wurde, zur Verfügung. Der Ablauf ist in Abbildung 7.5 zusammengefasst.

Der Test bestand aus zwei Phasen: einer kurzen Trainingsphase (zur Eingewöhnung) und einer ca. 12-minütigen Evaluierungsphase. Sowohl bei der Trainings- als auch bei der Evaluierungsphase musste der Proband eine Bewertung vornehmen. Bei Unregelmäßigkeiten bzw. auftretenden Unklarheiten während der Trainingsphase war diese zu wiederholen.

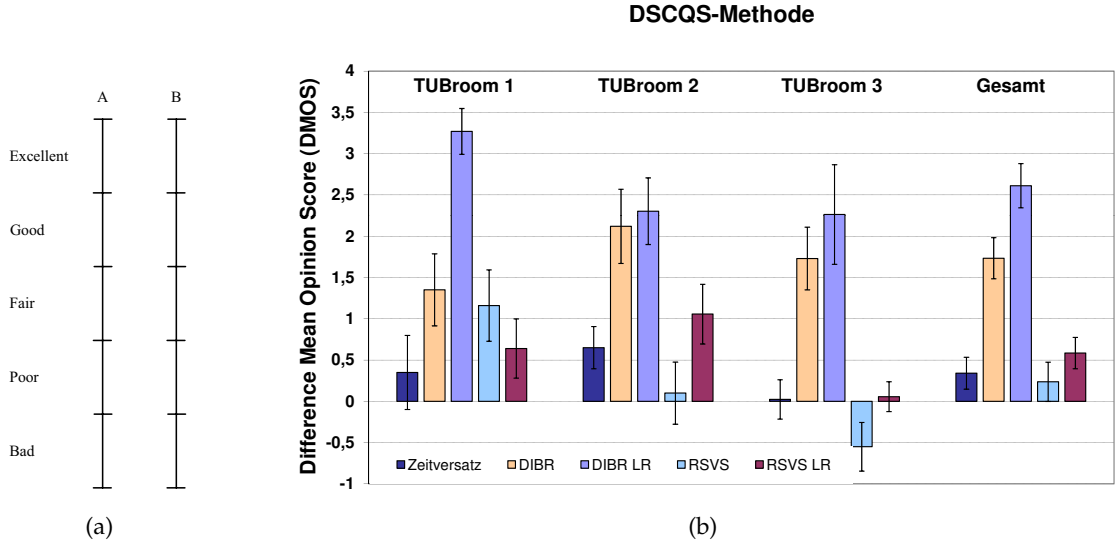
Als Bewertungskriterium sollte wie im Qualitätstest I der qualitative Gesamteindruck der gezeigten Testsequenzen beurteilt werden, um im voraus keine suggestive Beeinflussung der Probanden vorzunehmen.

7.3.2. Auswertung

Der durchschnittlich erzielte differentielle subjektive Qualitätseindruck (engl. *difference mean opinion score*, DMOS) und die Standardabweichung für drei Testsequenzen ist in Tabelle 7.2 dargestellt. Zur besseren Anschauung dient das Diagramm in Abbildung 7.6(b), welches ebenfalls den DMOS mit 95% Konfidenzintervall für alle Testsequenzen zeigt.

Tabelle 7.2.: Subjektiver Qualitätstest II (DSCQS-Methode): Durchschnittlicher DMOS und Standardabweichung [DMOS (± Std.Abw.)].

Datensatz	Zeitversatz	DIBR	DIBR LR	RSVS	RSVS LR
TUBroom 1	**0,35** (1,77)	1,35 (1,72)	3,27 (1,10)	1,16 (1,71)	0,64 (1,42)
TUBroom 2	0,65 (1,01)	2,12 (1,78)	2,30 (1,59)	**0,10** (1,49)	1,06 (1,43)
TUBroom 3	0,02 (0,94)	1,73 (1,50)	2,26 (2,38)	**-0,55** (1,16)	0,06 (0,72)
Gesamt	0,34 (1,32)	1,73 (1,70)	2,61 (1,83)	**0,24** (1,63)	0,58 (1,30)



(a) (b)

Abbildung 7.6.: Subjektiver Qualitätstest II (DSCQS-Methode): (a) Bewertungsskala, (b) Durchschnittlich erzielter subjektiver Qualitätseindruck (DMOS) mit 95% Konfidenzinterval.

Während des Tests ist eine Unregelmäßigkeit aufgetreten. Einer der Probanden hatte keine Bewertung während der Trainingsphase vorgenommen. Die Trainingsphase wurde demzufolge wiederholt.

Für alle drei Testsequenzen liefert RSVS mit einem DMOS von 0,24 den kleinsten Wert und somit die besten Ergebnisse. Die Probanden bewerteten somit sowohl die Referenzsequenzen als auch die mit RSVS konvertierten Testsequenzen nahezu gleich. Ebenfalls sehr gute Ergebnisse mit einem DMOS von 0,34 konnte die einfache Konvertierung mittels Zeitversatz erzielen. Allgemein schlechter bewerteten die Probanden jeweils die beiden Konvertierungsverfahren, bei denen sowohl die linke als auch die rechte Ansicht virtuell erzeugt wurde. DIBR schneidet generell am schlechtesten ab.

Da für die drei Testsequenzen jeweils eine spezielle Kameraführung gewählt wurde, müssen die Ergebnisse allerdings genauer spezifiziert werden.

TUBroom 1: Für die erste Testsequenz bekam „Zeitversatz“ mit einem DMOS von 0,35 die beste Bewertung. An zweiter Stelle war hier „RSVS LR“ mit einem DMOS von 0,64. Betrachtet man die Kameraführung, so kann man feststellen, dass sich die Kamera in der Mitte der Sequenz kurzzeitig vorwärts bewegt hat. Diese Kameraführung ist allerdings kritisch für RSVS, da die virtuelle Stereokamera zu weit vom Kamerapfad abweicht und somit verstärkt Transformationsfehler auftreten können. Durch die Synthese einer linken und rechten virtuellen Ansicht mit jeweils halbem Abstand zur Originalkamera wird diese Problematik abgeschwächt.

Die vorwärtsgerichtete Kameraführung hätte erwartungsgemäß auch negative Auswirkungen auf „Zeitversatz“ haben müssen. Die Begründung, dass bei „Zeitversatz“ zwei Originalansichten dargestellt werden und somit keine Einbußen bezüglich der Bildqualität vorlagen, kann hier allein nicht ausreichen. Bei genauer Betrachtung dieser Stereosequenz sind eine Reihe von 3D-Beeinträchtigungen zu erkennen. Zum einen lag wieder eine leichte vertikale Parallaxe vor, die bei längerer Betrachtung die Augenbelastung erhöht hätte. Zum anderen entstanden Trapezverzerrungen (engl. *keystone distortion*) [196], die vor allem bei konvergierendem Kamerasetup auftreten. Da bei einem einfachen Zeitversatz eine leichte Rotation der Kamera auftreten kann, sind solche Störungen nicht auszuschließen. Allerdings sind diese Effekte, die bei längerer Betrachtung eine Auswirkung auf das Tiefensehen besitzen, aufgrund der Kürze der Testsequenzen nicht von den Probanden beobachtet worden.

TUBroom 2: Bei der zweiten Testsequenz erhielt „RSVS“ mit einem DMOS von 0,1 die beste Bewertung. Hier hat sich die Kamera nach einer vorherigen Horizontalbewegung leicht nach oben und anschließend wieder nach unten bewegt. Da die Aufwärts- und Abwärtsbewegung gleichmäßig und langsam war, ist die virtuelle Kamera bei „RSVS“ nur geringfügig vom Kamerapfad abgewichen, wodurch Transformationsfehler nicht sichtbar waren. Aus dem gleichen Grund wurde „Zeitversatz“ mit einem DMOS von 0,65 recht gut bewertet. Allerdings trat auch hier durch die Aufwärtsbewegung eine vertikale Parallaxe auf, die nicht von den Probanden realisiert wurde.

TUBroom 3: Bei der dritten Testsequenz bewegte sich die Kamera fast vollständig seitwärts, so dass „RSVS“ und „Zeitversatz“ mit einem DMOS von -0,55 und 0,02 besonders gut abschnitten. Die Besonderheit bei dieser Testsequenz lag in zwei Eigenschaften. Zum einen „wanderte“ das Vordergrundobjekt (der Würfel) seitwärts in das Bild hinein, so dass es kurzzeitig vom Bildrand verdeckt wurde. Da bei der Referenzsequenz ein paralleles Kamerasetup (ohne Shift-Sensor) verwendet worden ist, erschien das Vordergrundobjekt durch rein negative Parallaxe vor dem Bildschirm. Die teilweise Verdeckung durch die Begrenzung des Bildschirms vermittelt allerdings den Eindruck, dass das Objekt

hinter dem Bildschirm sein muss. In diesem Fall hatte somit die Referenzsequenz einen „Bruch“ zwischen monokularer und stereoskopischer Tiefenwahrnehmung (engl. *break in depth cues*) [109], weshalb „RSVS“ mit dem Shift-Sensor und damit der Wahl der Null-Parallaxen-Ebene weiter im Vordergrund eine bessere Bewertung als die Referenzsequenz erhielt.

Durch die fast vollständig gleichmäßige Seitwärtsbewegung der Kamera liefert „Zeitversatz“ fast identische stereoskopische Bilder wie die Referenzsequenz, was sich an dem geringen DMOS von 0,02 ablesen lässt. Die Probanden haben kaum einen Unterschied zwischen Test- und Referenzsequenz entdecken können.

Die zweite Eigenschaft der Sequenz ist wiederum mit dem Würfel im Vordergrund verbunden. Durch die Seitwärtsbewegung befindet sich die Kamera in der zweiten Hälfte der Sequenz an einer Position, von der aus der Würfel direkt vor der geöffneten Tür aufgenommen wird (siehe Abbildung 7.4(c)). An den oberen Kanten des Würfels treten somit große Tiefensprünge auf, die die Aufdeckungproblematik bei DIBR besonders zum Vorschein bringt. Zwar wurde für „DIBR“ und „DIBR LR“ jeweils eine Vorfilterung der Tiefenkarten durchgeführt, so dass diese Problematik umgangen wird, allerdings wird dadurch der Vordergrund stärker mit dem Hintergrund „verschmiert“, wie in Kapitel 6 bereits erörtert wurde (siehe dazu ebenfalls Abbildung 6.6(d)). Diese Eigenschaft macht sich in der schlechten Bewertung der Testsequenz für „DIBR“ (DMOS von 1,73) und „DIBR LR“ (DMOS von 2,26) bemerkbar.

Auf die beiden Fragen („Ist Ihnen bei dem Test irgendetwas Besonderes aufgefallen?“ und „Welche Störungen bzw. Fehler sind Ihnen aufgefallen?“), die am Ende des Tests gestellt wurden, antworteten drei der Probanden, dass die gezeigten Sequenzen allgemein etwas zu kurz waren. Weitere Fehler und Störungen, die den Probanden aufgefallen sind, waren: Spiegelungsfehler (6 Probanden), die bereits beim Rendern durch die hochfrequenten Bereiche entstanden sind, Geisterbilder (6 Probanden) durch die schlechte Kanaltrennung, Veränderungen der Würfelform (5 Probanden) bei DIBR und Transformationsfehler[4] (2 Probanden), die vor allem bei RSVS in der Sequenz *TUBroom 1* auftraten.

7.4. Kapitelzusammenfassung und Diskussion

In diesem Kapitel wurde eine umfassende Untersuchung des neuen Stereosynteseverfahrens RSVS mit subjektiven Qualitätstests durchgeführt. Dabei waren die Testbedingungen standardkonform mit den ITU-Empfehlungen. In einem ersten Test wurde RSVS über die SS-Methode qualitativ mit einem einfachen 3D-Konvertierungsverfahren („Zeitversatz“)

[4] Da die beiden Probanden fachfremd waren, verwendeten sie eine eigene bildliche Wortwahl.

und dem ursprünglichen 2D-Format verglichen. Der Test umfasste 25 Testsequenzen (18 von der BBC-Dokumentation *Planet Erde* und 7 Handkamerasequenzen).

Im zweiten Qualitätstest wurde die DSCQS-Methode verwendet, da drei synthetische stereoskopische Referenzsequenzen vorlagen. Hier wurde ein Vergleich zwischen den 3D-Syntheseverfahren „Zeitversatz“, „RSVS“, „RSVS LR“, „DIBR“ und „DIBR LR“ vorgenommen. Beide Tests zeigten, dass RSVS hervorragende Konvertierungseigenschaften besitzt und qualitativ hochwertige stereoskopische Sequenzen erzeugt. Allerdings sei darauf hingewiesen, dass das gesamte Testmaterial Sequenzen enthielt, welche mittels RSVS konvertierbar waren. DIBR, das im Vergleich mit RSVS weitaus schlechter abgeschnitten hat, lässt sich für alle Arten von Videosequenzen einsetzen, wenn dichte Tiefenkarten vonhanden sind, und ist somit sehr viel flexibler als RSVS. Dennoch lässt sich festhalten, dass RSVS für spezielle Sequenzen (statische Szene, bewegte Kamera) bessere Konvertierungsergebnisse als DIBR liefert.

„Zeitversatz“ als mögliche Konvertierungsform wurde ebenfalls qualitativ hochwertig eingestuft. Im ersten Test sind allerdings deutlich die stärkeren Einschränkungen bezüglich Kameraführung (Handkamerasequenzen) aufgezeigt worden. Generell muss aber an dieser Stelle deutlich gemacht werden, dass keiner der beiden Tests ausreichend war, um die Auswirkungen von 3D-Artefakten wie vertikale Parallaxe oder Trapezverzerrungen zu überprüfen. Diese Fehler sind bei „Zeitversatz“ zwar aufgetreten, wurden in der Kürze der Sequenzen allerdings von keinem der Probanden bemerkt. Die Auswirkungen wie erhöhte Augenbelastung bzw. - bei längerer Betrachtung - Kopfschmerzen sind bereits hinlänglich bekannt.

Langzeittests, bei denen Probanden längere Zeit mit diesen Artefakten konfrontiert werden, könnten das bestätigen. Jedoch wird dafür zum einen ein größerer Testdatensatz benötigt und zum anderen sollte dazu ein Testablauf standardisiert werden. Subjektive Qualitätstest sollten in der Regel nicht länger als 30 Minuten pro Proband dauern. Zudem könnten Probanden zwei verschiedene Konvertierungsverfahren nicht unbedingt direkt nacheinander bewerten. Zeigt man einem Probanden z. B. eine längere Testsequenz mit vertikaler Parallaxe, so dass dieser davon Kopfschmerzen bekommt, und direkt im Anschluss die gleiche Testsequenz ohne vertikale Parallaxe, so ist davon auszugehen, dass der Proband seine Kopfschmerzen behält und damit die zweite Sequenz entweder gleich oder vielleicht sogar schlechter bewerten würde.

8. Einführung in den Umgang mit dynamischen Szenen

In den voran gegangenen Kapiteln wurde detailliert die Analyse und Rekonstruktion statischer Szenen, die mit einer bewegten Kamera aufgenommen wurden, mit dem Ziel der Synthese stereoskopischer Ansichten beschrieben. Für die Rekonstruktion der 3D-Struktur und der Kameraparameter wurde SfM verwendet, das nur auf eine einzelne Relativbewegung (bewegte Kamera $\leftrightarrow$ statische Szene oder statische Kamera $\leftrightarrow$ bewegte Szene) anwendbar ist. In diesem Kapitel soll nun auf die Problematik der 3D-Rekonstruktion eingegangen werden, wenn unabhängig von der Kamerabewegung bewegte Objekte (engl. *independently moving object*, IMO[1]) in der Szene auftreten, d.h. die aufgenommene Szene nun mehr *dynamisch* ist. Die geometrischen Eigenschaften, wie die Epipolargeometrie, sind bei dynamischen Szenen nicht mehr trivial [157, 195].

Die Erweiterung von SfM auf dynamische Szenen ist ein noch recht junges Forschungsfeld mit ersten vielversprechenden Segmentierungsergebnissen für Merkmalspunkte [49, 157, 192, 192, 195] und einigen wenigen 3D-Rekonstruktionsergebnissen der Szene mit den entsprechenden IMOs [54, 76, 77, 99, 178]. Dieses Kapitel soll hierzu lediglich eine Einführung in den Umgang mit dynamischen Szenen geben, auf die zusätzlichen Probleme eingehen und gleichzeitig einen Ausblick liefern, um die Einschränkungen von SfM und damit auch von RSVS zu reduzieren.

8.1. Problemstellung und Überblick

Der eigentliche SfM-Ansatz muss für dynamische Szenen um die zusätzlichen IMOs erweitert werden. Diese Erweiterung wird im Allgemeinen auch als *multi-body SfM* (engl., SfM für mehrere Objekte, MSfM) bezeichnet [49, 157]. Das MSfM-Problem lässt sich durch den folgenden Satz beschreiben:

Satz 8.1.1 *Seien N Merkmalspunkte, die sowohl zu einem statischen Hintergrund als auch zu K IMOs gehören, L Kameraansichten und die Korrespondenzinformation der Merkmalspunkte zwischen den Ansichten gegeben, dann können daraus die 3D-Koordinaten der Merkmalspunkte und die externen Kameraparameter bestimmt werden.*

[1] Ein IMO ist ein in sich selbst statisches Objekt, das seine Form über die Zeit nicht ändert. Eine Person, die sich innerhalb der Szene bewegt, würde demnach aus mehreren IMOs bestehen (z. B. Füße, Hände, Unterschenkel, etc.).

Das Problem kann durch Segmentierung der Merkmalspunkte, die mittels LK-Tracker (siehe Abschnitt 5.2) über die Sequenz verfolgt wurden, in mehrere statische SfM-Probleme zerlegt werden, d.h. sind die Merkmalspunkte entsprechend ihrer Zugehörigkeit zum Hintergrund und zu den IMOs klassifiziert, können diese einzelnen Datensätze verwendet werden, um jeweils eine SfM-Analyse durchzuführen. In diesem Kapitel liegt somit der Fokus in der Segmentierung des Datensatzes an Merkmalspunkten in Hintergrund und IMOs, damit im Anschluss jeweils separat mit den segmentierten Merkmalspunkten eine 3D-Rekonstruktion durchgeführt werden kann.

Die Segmentierungsverfahren, die sich mit dem MSfM-Problem befassen, können in drei Kategorien aufgeteilt werden: Verfahren, die auf dem optischen Fluss basieren [78, 114], gehen von einer Szene aus, die aus mehreren Ebenen mit variierender Tiefe besteht. Unter dieser einfachen Annahme geben die verschiedenen Werte des optischen Flusses an, welche Szenenbereiche zusammen gehören und somit entsprechend segmentiert werden können. Problematisch bei diesen Ansätzen ist der Umgang mit nahen Objekten, die ebenfalls zum statischen Hintergrund gehören. Diese werden i.d.R. mitsegmentiert und einer unabhängigen Bewegung zugeordnet. Damit verschlechtert sich allerdings die Rekonstruktion mittels SfM, da zusätzliche Bereiche der Szene, die die Rekonstruktionsergebnisse positiv beeinflussen können, nicht berücksichtigt werden.

Statistische Methoden gehören zur zweiten Kategorie. In [54] wird SIS (*sequential importance sampling*) angewandt, um gleichzeitig SfM für mehrere IMOs durchzuführen. Die empirische Verteilung der Objektbewegung und der Parameter zur Segmentierung der Merkmalspunkte wird a posteriori über Gewichtungsparameter geschätzt.

Zur dritten Kategorie gehören Verfahren, die die Epipolargeometrie ausnutzen. Dabei wird die Annahme getroffen, dass bewegte Objekte in der Szene starr sind und nicht ihre Form ändern. Der am meisten verbreitete Ansatz ist, die individuellen Fundamentalmatrizen für jede unabhängige Bewegung in der aufgenommenen Szene zu schätzen, um die Merkmalspunkte bezüglich der Epipolarbedingung zu klassifizieren [49, 99, 192, 195]. In [182] wird GRIC verwendet, um sowohl den Datensatz zu segmentieren als auch für jeden segmentierten Teildatensatz das entsprechend zugrunde liegende Modell zu bestimmen. Ein weiterer Ansatz wird in [25] vorgeschlagen, welcher eine Erweiterung der Faktorisierung für SfM darstellt.

Das neue Verfahren zur Segmentierung und 3D-Rekonstruktion der mittels LK-Tracker verfolgten Merkmalspunkte, welches hier vorgestellt wird, gehört zur dritten Kategorie und ist somit ein geometrischer Segmentierungsansatz, der auf der Epipolarbedingung basiert. Die Schritte zur 3D-Rekonstruktion der einzelnen unabhängigen Bewegungen können wie folgt zusammengefasst werden:

1. Detektion von Merkmalspunkten mit dem Harris-Detektor (siehe Abschnitt 4.2.1)

2. Verfolgung von Merkmalspunkten mit dem pyramidenförmigen LK-Tracker (siehe Abschnitt 5.2.4)
3. Detektion von Schlüsselbildern mit Hilfe des GRIC-Ansatzes (siehe Abschnitt 5.3)
4. Segmentierung der Merkmalspunkte über die Epipolarbedingung
5. Separate 3D-Rekonstruktion des Hintergrunds und der IMOs (siehe Abschnitte 5.4 und 5.5)

Die geometrische Segmentierung des Datensatzes in Merkmalspunkte des Hintergrunds und Merkmalspunkte der IMOs, welche sich der Epipolarbedingung bedient, wird mit Hilfe robuster Schätzverfahren durchgeführt. Als robustes Schätzverfahren wird der in Abschnitt 3.3.1 beschriebene RANSAC-Algorithmus verwendet.

8.2. Ansatz zur 3D-Rekonstruktion dynamischer Szenen

8.2.1. Tracking und Detektion von Schlüsselbildern

Die Verfolgung von Merkmalspunkten in einer Videosequenz und die Detektion von Schlüsselbildern mit Hilfe des GRIC-Ansatzes wurden in den Abschnitten 5.2.4 und 5.3 hinreichend beschrieben. Bei dem Umgang mit dynamischen Szenen sind allerdings einige Eigenschaften zu beobachten, die einer Modifikation der Algorithmen und einer näheren Untersuchung bedürfen.

Zum einen wird eine genauere Definition der verfolgten Merkmalspunkte bzw. deren Trajektorien benötigt, da diejenigen Merkmalspunkte, die nur über einige wenige Bilder verfolgt werden konnten, ggf. nicht geometrisch segmentiert werden können. Zum anderen ist die Detektion von Schlüsselbildern bei dynamischen Szenen kritischer zu betrachten, da durch die IMOs und damit die zusätzliche Bewegung der GRIC-Wert schneller negativ werden kann (d.h. frühzeitiger Schlüsselbilder detektiert werden).

In Gl. 5.37 wurden drei Regelungsparameter (λ_1, λ_2 und λ_3) verwendet, um die GRIC-Werte für die beiden zugrunde liegenden Modelle (Homographie H und Epipolargeometrie F) zu steuern. Der Parameter λ_1 diente dabei als eine Art „Bestrafung“ für das höhere Modell (Epipolargeometrie) und wurde in [185] mit $\lambda_1 = \log 4$ vorgeschlagen. In [181] und [183] hat Torr für λ_1 einen Wert zwischen 1 und 2 angegeben, je nach Größe des Datensatzes, d.h. $\lambda_1 = 2$ für kleinere Datensätze. Da die unterschiedlichen Bewegungen in der Szene einen größeren Einfluss auf den Transformationsfehler bei der Berechnung der Homographie besitzen, sollte dieser durch Wahl eines größeren Wertes für λ_1 ausgeglichen werden, um die Basislänge zwischen Schlüsselbildern bzw. deren Kamerapositionen zu erhöhen. Abbildung 8.1 zeigt die GRIC-Verläufe für $\lambda_1 = \log 4$ [185] und $\lambda_1 = 2$ [183] für dynamische und statische Szenen.

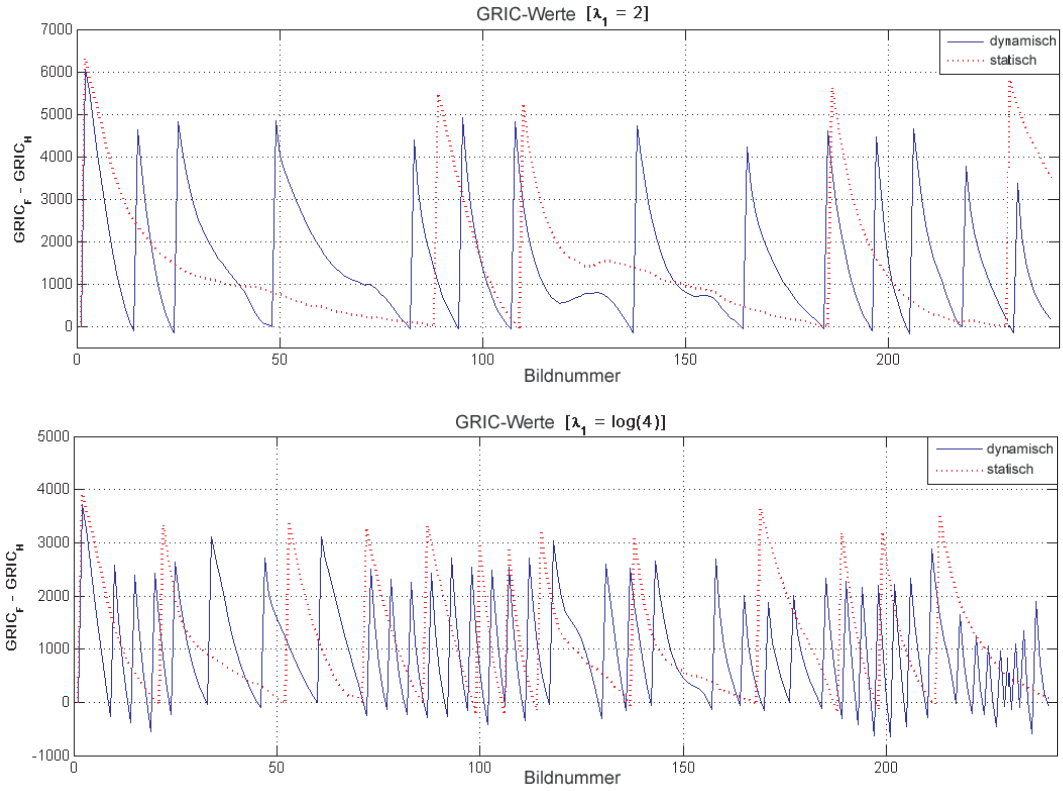


Abbildung 8.1.: Vergleich der GRIC-Verläufe bei statischen und dynamischen Szenen für verschiedene Werte von λ_1.

Für die Berechnung der GRIC-Verläufe wurden zwei fast identische synthetische Sequenzen verwendet: *TUBroom-IMO* und *TUBroom*. Beide Sequenzen wurden mit den gleichen Kameraeinstellungen und dem gleichen Kamerapfad gerendert und haben eine Länge von 240 Einzelbildern. Der statische Hintergrund ist in beiden Fällen ebenfalls gleich. Bei *TUBroom-IMO* wurden zusätzlich zwei IMOs in die Szene integriert (siehe Abbildung 8.4), die sich unabhängig von der Kamera bewegen, so dass die Szene dynamisch ist.

Die Abbildung 8.1 macht deutlich, dass die GRIC-Werte für die dynamische Szene schneller negativ und somit eher Schlüsselbilder detektiert werden. Für $\lambda_1 = \log 4$ (untere Abbildung) werden insgesamt 41 Schlüsselbilder bei der dynamischen Szene detektiert (d.h. durchschnittlich nach 5,85 Einzelbildern). Diese Anzahl ist allerdings recht groß bei einer Sequenz mit 240 Einzelbildern. Der Kameraabstand zwischen den Schlüsselbildern ist mit großer Sicherheit äußerst gering. Bei der statischen Szene werden lediglich 14 Schlüsselbilder detektiert (d.h. durchschnittlich nach 17,14 Einzelbildern).

Für $\lambda_1 = 2$ (obere Abbildung) hat sich die Anzahl der Schlüsselbilder bei der dynamischen Szene auf 14 reduziert, d.h. durchschnittlich nach 17,14 Einzelbildern wird ein Schlüsselbild detektiert. Da sowohl für die geometrische Segmentierung der Merkmalspunkte als auch für die anschließende Rekonstruktion eine große Basislänge zwischen

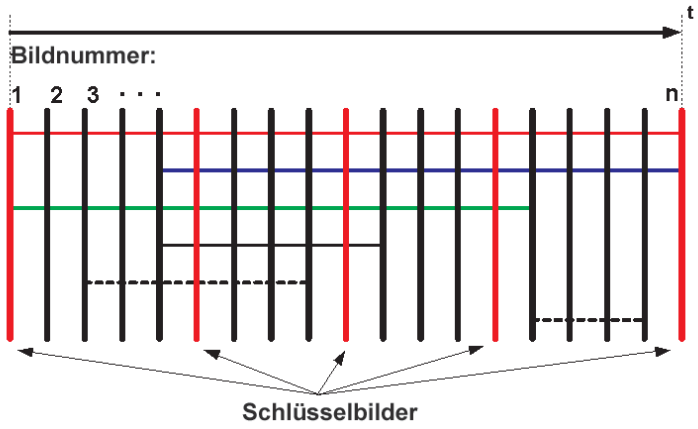


Abbildung 8.2.: Klassifikation von Trajektorienarten

den Kamerapositionen gefordert wird, sollte für dynamische Szenen der Regelungsparameter auf $\lambda_1 = 2$ erhöht werden.

8.2.2. Geometrische Segmentierung

Nachdem die Schlüsselbilder detektiert worden sind, werden die Merkmalspunkttrajektorien über geometrische Zusammenhänge segmentiert. Abbildung 8.2 zeigt schematisch n Einzelbilder einer Videosequenz mit einigen detektierten Schlüsselbildern (senkrechte rote Linien) und fünf verschiedenen Arten an Merkmalspunkttrajektorien. Für die anschließende Segmentierung der Merkmalspunkte und die 3D-Rekonstuktion werden nur diejenigen in Betracht gezogen, die mindestens in zwei aufeinander folgenden Schlüsselbildern sichtbar sind (d.h. die oberen vier horizontalen Linien).

Für jede unabhängige Bewegung innerhalb der Sequenz existiert ein geometrischer Zusammenhang über die Epipolargeometrie, d.h. zu jeder Relativbewegung zwischen Kamera und Objekt (Hintergrund oder IMO) gibt es eine zugehörige Fundamentalmatrix, F_i, die die Epipolarbedingung (Gl. 2.26) erfüllt. Die zugehörigen Fundamentalmatrizen lassen sich durch robuste Schätzverfahren, wie sie in Kapitel 3 vorgestellt wurden, ermitteln, indem mittels RANSAC die Merkmalspunkte in einen Satz korrekter Zuordnungen, die die Epipolarbedingung für die Matrix F_1 erfüllen, und Ausreißer zerlegt werden. Die korrekten Zuordnungen gehören dabei zu der dominanten Bewegung (i.d.R. die Bewegung zwischen Kamera und statischem Hintergrund).

Für die Ausreißer gilt demnach, dass sie entweder falsche oder schlechte Zuordnungen sind oder zu den restlichen Bewegungen in der Szene gehören (i.d.R. die Bewegung zwischen der Kamera und den IMOs). Wendet man RANSAC erneut auf den Satz an Ausreißern an, werden einige der Zuordnungen die Epipolarbedingung bezüglich einer neuen Fundamentalmatrix F_2 erfüllen, die der Bewegung eines IMOs entspricht. Diese Prozedur kann wiederholt werden, solange glaubwürdige Fundamentalmatrizen bestimmt

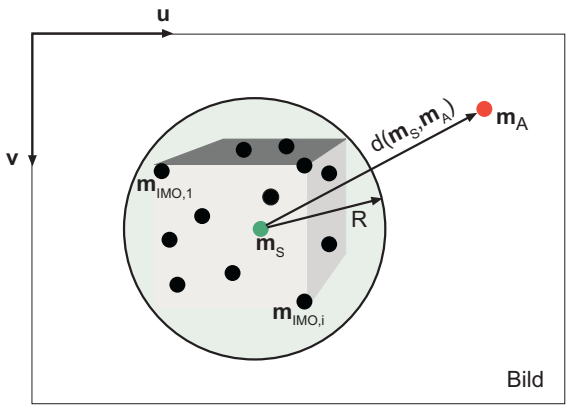


Abbildung 8.3.: Robuste IMO-Segmentierung

werden können ($F_3, F_4, \ldots, F_{K+1}$). Die Glaubwürdigkeit einer verlässlichen Fundamentalmatrix hängt dabei sowohl von der Relativbewegung zwischen der Kamera und einem IMO als auch von der Anzahl an Merkmalspunkten auf dem IMO ab.

Durch die Ausführung von RANSAC und re-RANSAC[2] lassen sich somit die Merkmalspunkttrajektorien des Hintergrunds bzw. der IMOs über alle Schlüsselbilder klassifizieren.

Erhöhung der Robustheit

Bei der Segmentierung des Datensatzes in Merkmalspunkttrajektorien des Hintergrunds und der IMOs werden hohe Ansprüche an den RANSAC-Algorithmus gestellt. Zwar ist RANSAC dafür bekannt, dass er in der Lage ist, mit einem Ausreißeranteil von mehr als 50% umzugehen, dieser kann aber z. B. bei re-RANSAC zur Klassifizierung der IMOs weit überschritten werden, wenn mehrere IMOs in der Szene vorhanden sind und zusätzlich der Datensatz stark verrauscht ist. Zudem kann die Relativbewegung zwischen der Kamera und einem IMO ähnlich der Relativbewegung des Hintergrunds sein. Bei der Klassifizierung der Merkmalspunkte des Hintergrunds könnten somit aufgrund von Rauscheinflüssen einige Punkte des Hintergrunds als Ausreißer klassifiziert werden, die in Teilen ggf. die Epipolarbedingung für ein IMO erfüllen und diesem zugeordnet werden.

Um dieses möglichst zu verhindern bzw. stärker zu unterdrücken, wird ein maximaler Radius R um den Schwerpunkt der segmentierten Merkmalspunkte eines IMOs definiert, innerhalb dessen die Merkmalspunkte tatsächlich dem IMO zugeordnet werden:

$$\mathbf{m}_i = \begin{cases} \mathbf{m}_{IMO,i} & , wenn\, d(\mathbf{m}_i, \mathbf{m}_S) \leq R \\ \mathbf{m}_A & , sonst \end{cases} \tag{8.1}$$

[2]Wiederholung von RANSAC auf den jeweiligen Satz an Ausreißern.

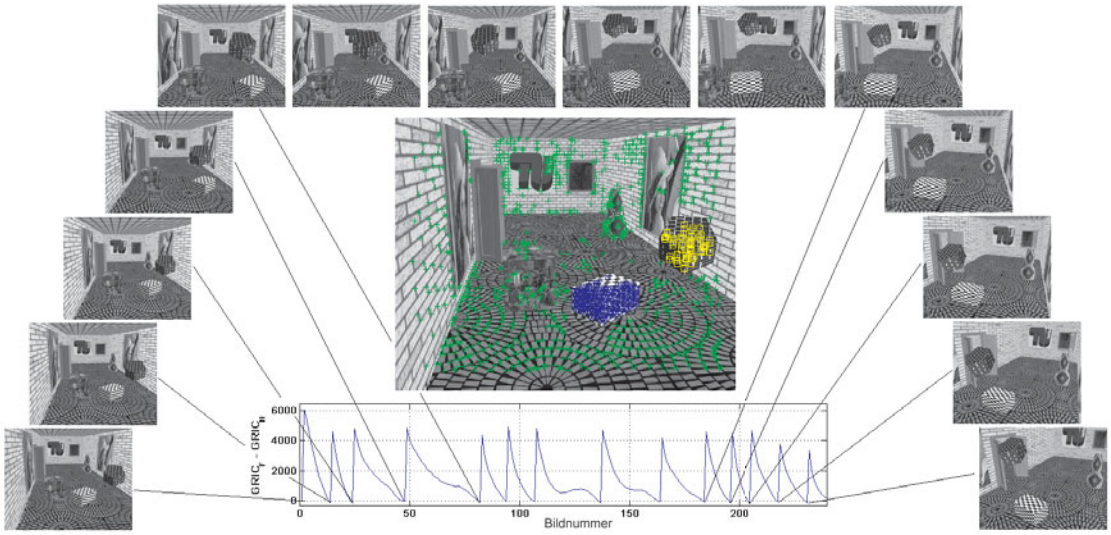


Abbildung 8.4.: Klassifizierung der Merkmalspunkte, GRIC-Verlauf und detektierte Schlüsselbilder der Sequenz *TUBroom-IMO*: Hintergrund (grün), IMO 1 (blau), IMO 2 (gelb)

mit

$$R = \frac{\sum_{i=1}^{N} d(\mathbf{m}_i, \mathbf{m}_S)}{N} + \nu \cdot \sigma \,, \tag{8.2}$$

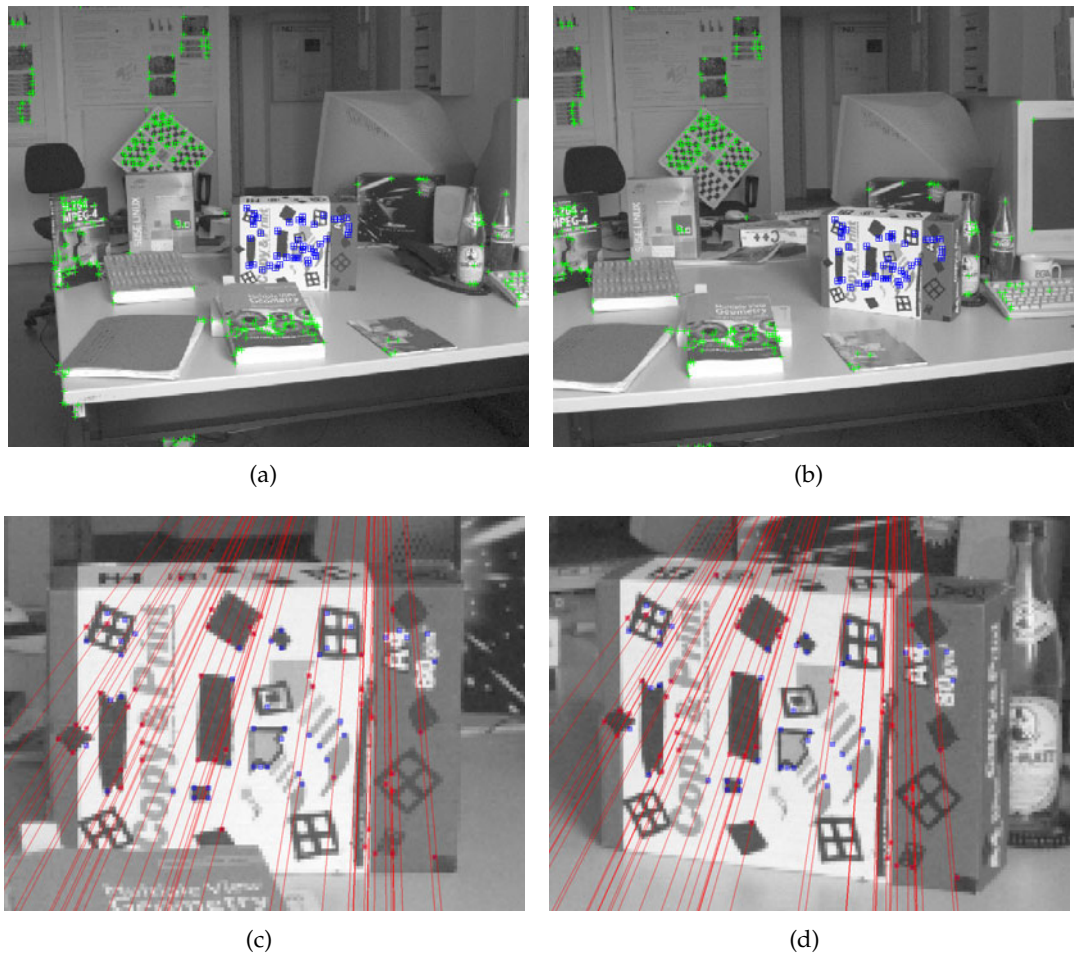
wobei $d(\mathbf{m}_i, \mathbf{m}_S)$ der Euklidische Abstand zwischen einem Merkmalspunkt $\mathbf{m}_i$ und dem Schwerpunkt $\mathbf{m}_S$ ist, $\mathbf{m}_{IMO,i}$ ein Merkmalspunkt, der dem IMO zugeordnet wird, und $\mathbf{m}_A$ ein Ausreißer ist, der weder dem Hintergrund noch einem IMO zugeordnet wird. σ ist die Standardabweichung der Abstände und ν ein Gewichtungsfaktor, der den Radius R reguliert.

Abbildung 8.3 zeigt ein Beispiel für die robuste Segmentierung. Bei dem re-RANSAC werden Merkmalspunkte sowohl auf dem IMO als auch im anderen Bereich des Bildes klassifiziert. Durch die Definition eines Toleranzbereichs (der Radius R um den Schwerpunkt $\mathbf{m}_S$) werden nur diejenigen Merkmalspunkte dem IMO zugeordnet, die tatsächlich auf diesem liegen. Der Merkmalpunkt $\mathbf{m}_A$ wird als endgültiger Ausreißer klassifiziert.

In Abbildung 8.4 sind die 14 detektierten Schlüsselbilder, der GRIC-Verlauf und das erste Bild der synthetischen Sequenz *TUBroom-IMO* mit den klassifizierten Merkmalspunkten dargestellt.

Guided Matching

Nachdem die Merkmalspunkttrajektorien entsprechend ihrer Bewegungen den Objekten zugeordnet sind, stellt die Rekonstruktion zumindest der IMOs dann ein Problem



(a) (b)

(c) (d)

Abbildung 8.5.: Klassifizierung der Merkmalspunkte (grün: Hintergrund, blau: IMO) der Sequenz *Box*: (a) Bild 1 und (b) Bild 13; (c) - (d) guided-Matching entlang der Epipolarlinien.

dar, wenn nicht genügend Merkmalspunkte auf diesen liegen und somit wenig Punktkorrespondenzen für die Triangulation bereitstehen. Eine Möglichkeit für die Erhöhung der Punktkorrespondenzen ist das in Abschnitt 4.3.2 vorgestellte *guided-Matching*. Da die Fundamentalmatrizen zwischen den Schlüsselbildern für jede unabhängige Bewegung bekannt sind, können diese verwendet werden, um weitere Merkmalspunkte, die in jedem Schlüsselbild neu detektiert werden, entlang der Epipolarlinien zuzuordnen.

Abbildung 8.5 zeigt die Ergebnisse der Segmentierung und des guided-Matchings zwischen den Bildern 1 und 13 der Stop-Motion-Sequenz *Box* mit insgesamt 13 Einzelbildern. Die Merkmalspunkte des Hintergrundes sind grün und die des IMOs blau dargestellt. In den Abbildungen 8.5(c) und 8.5(d) sind jeweils die Epipolarlinien und zusätzlich ermittelte Punktkorrespondenzen (rot) dargestellt, die ebenfalls dem IMO zugeordnet werden. Als weitere Einschränkung des Suchbereichs bei guided-Matching können die bereits bekannten Bewegungsvektoren der klassifizierten Merkmalspunkte verwendet werden.

Ziel: Aufteilung des Datensatzes von Merkmalspunkten in Hintergrund und IMOs.

1. Berechne die Fundamentalmatrix zwischen dem ersten und zweiten Schlüsselbild mit Hilfe des RANSAC-Algorithmus und weise die korrekten Zuordnungen von Merkmalspunkten dem Hintergrund zu.
2. Berechne die Fundamentalmatrix für die Ausreißer aus dem vorherigen Berechnungsschritt erneut mittels RANSAC und weise die korrekten Zuordnungen einem IMO zu.
3. Bestimme den Schwerpunkt der korrekten Zuordnungen aus Berechnungsschritt 2 und berechne den Abstand zwischen Schwerpunkt und korrekten Zuordnungen. Ist der Abstand größer als der Radius R (Gl. 8.1), entferne die entsprechenden Merkmalspunkte.
4. Erhöhe die Anzahl an Merkmalspunkten auf dem IMO in den beiden Schlüsselbildern durch guided-Matching.
5. Wiederhole die Berechnungsschritte 2 bis 4 solange die Schätzung der Fundamentalmatrix zulässig ist.
6. Gehe zu dem nächsten Schlüsselbild und berechne die Fundamentalmatrix zwischen dem letzten und dem aktuellen Schlüsselbild mit den klassifizierten Merkmalspunkten und weise neue Merkmalspunkte dem Hintergrund bzw. den IMOs durch Wiederholung der Berechnungsschritte 2 bis 5 zu.
7. Wiederhole Berechnungsschritt 6 für alle Schlüsselbilder der Sequenz.

Ergebnis: Klassifizierte Merkmalspunkte entsprechend ihrer Zugehörigkeit zum Hintergrund oder einem IMO.

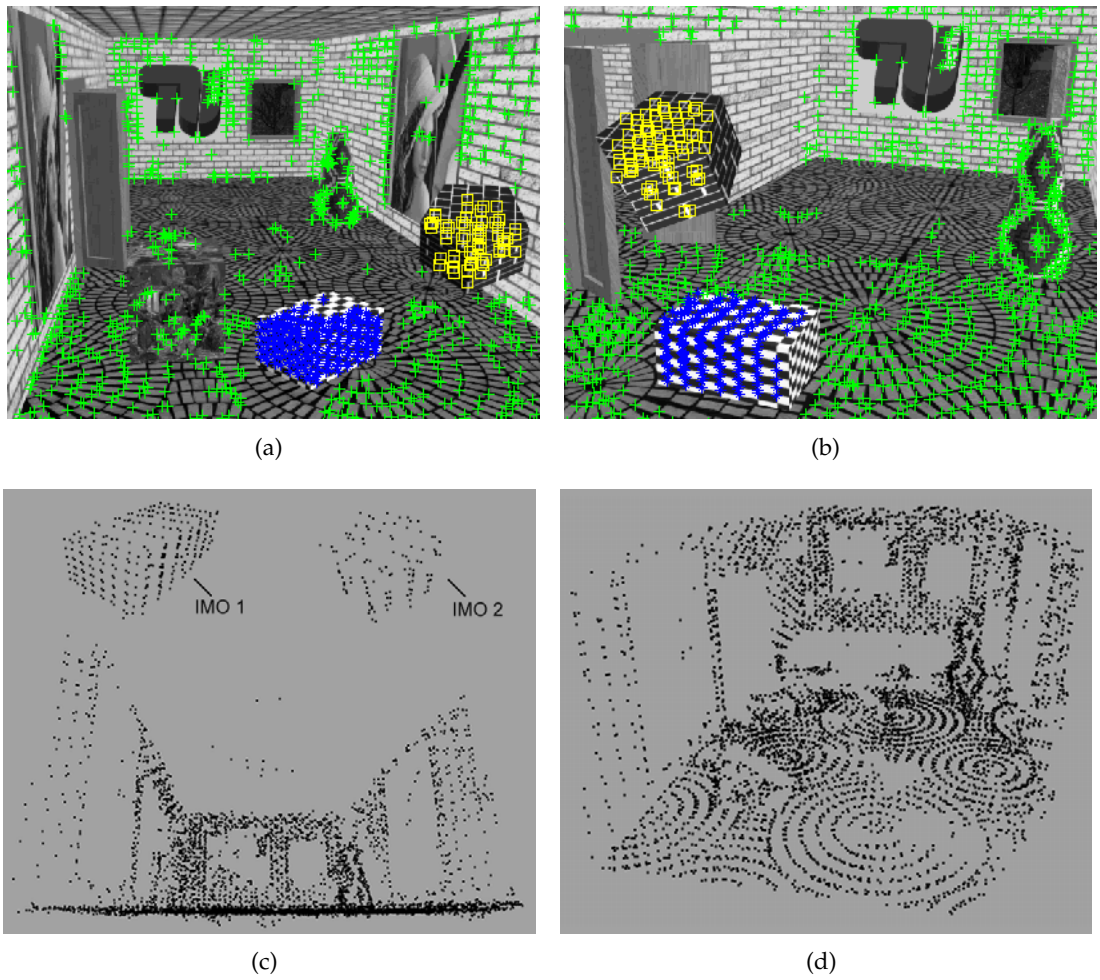
Algorithmus 11: Zusammenfassung der Berechnungsschritte zur Segmentierung von Merkmalspunkttrajektorien

Der gesamte Ablauf der Segmentierung der Merkmalspunkte entsprechend ihrer Relativbewegung ist in Algorithmus 11 zusammengefasst.

8.3. Rekonstruktionsergebnisse

In Abbildung 8.6 sind die Rekonstruktionsergebnisse der synthetischen Sequenz *TUBroom-IMO* dargestellt. Die oberen beiden Bilder zeigen die Segmentierungsergebnisse der Merkmalspunkte des Hintergrunds (grün) und der beiden IMOs (blau und gelb). In den unteren beiden Bildern sind die einzelnen 3D-Rekonstruktionen des Hintergrunds (Abbildungen 8.6(c) unten und 8.6(d)) und der beiden IMOs (Abbildungen 8.6(c) oben) dargestellt.

Die Rekonstruktion des IMOs aus der *Box*-Sequenz war auch mit guided-Matching nicht möglich, da fast alle Merkmalspunkte auf der Vorderseite der Box lagen und somit auf einer Ebene (siehe Abbildung 8.5). Für die Sequenz *Sofa*, eine Stop-Motion-Sequenz



(a) (b)

(c) (d)

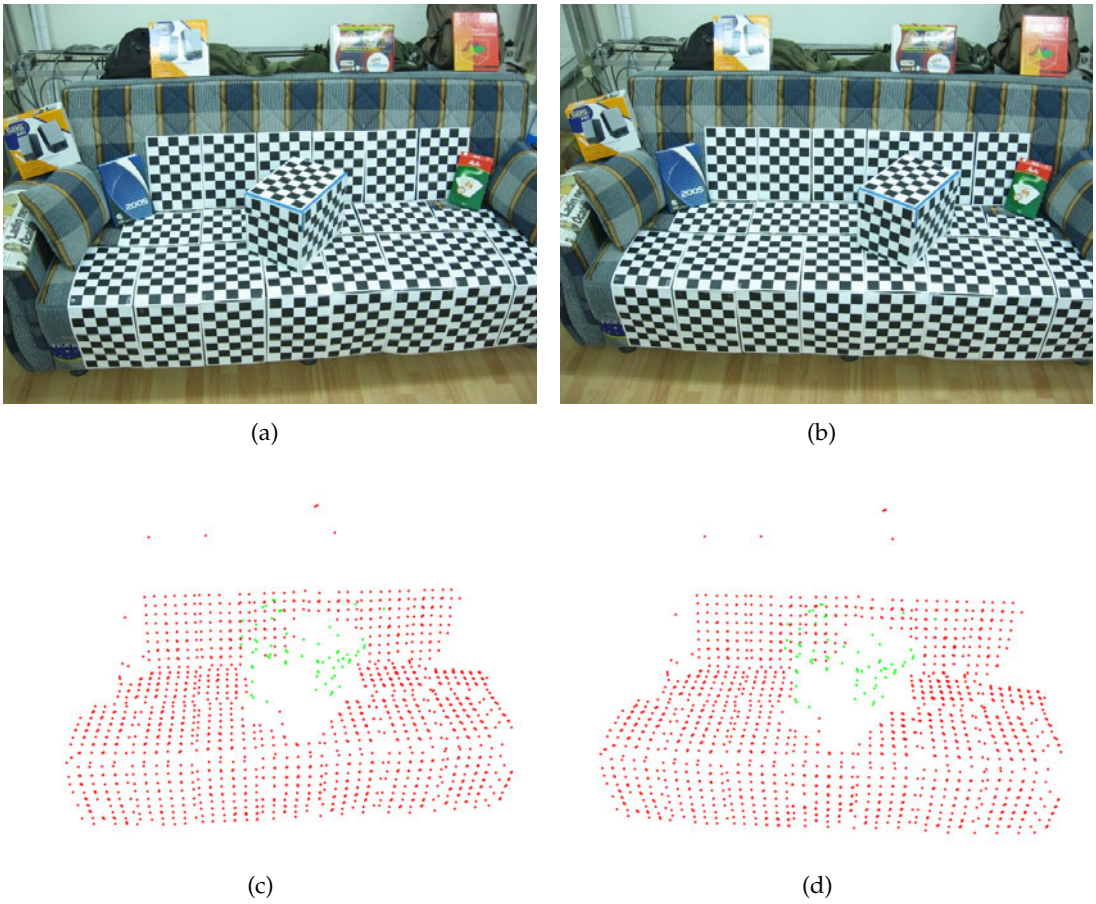
Abbildung 8.6.: Klassifizierung der Merkmalspunkte (grün: Hintergrund, blau bzw. gelb: IMOs) der Sequenz *TUBroom-IMO*: (a) Bild 1 und (b) Bild 201; 3D-Rekonstruktionen: (c) einzeln rekonstruierte IMOs (oben) und Vorderansicht des rekonstruierten Hintergrunds (unten), (d) rekonstruierter Hintergrund.

mit ebenfalls 13 Einzelbildern, konnte sowohl der Hintergrund als auch das IMO rekonstruiert werden. Abbildung 8.7 zeigt das erste und letzte Bild der Sequenz und die 3D-Punktwolken aus den entsprechenden Perspektiven der beiden Aufnahmen. Das separat rekonstruierte IMO (grün) wurde nachträglich in den rekonstruierten Hintergrund integriert[3].

8.4. Kapitelzusammenfassung, Diskussion und Ausblick

Zusammenfassung: In diesem Kapitel wurde eine Einführung in den Umgang mit der Rekonstruktion dynamischer Szenen geliefert. Dabei wurde ein neues Verfahren vorge-

[3]Das IMO wurde entsprechend der Position zum Zeitpunkt der ersten Kameraaufnahme in die Hintergrundrekonstruktion integriert.



(a) (b)

(c) (d)

Abbildung 8.7.: 3D-Rekonstruktion der *Sofa*-Sequenz: (a) Bild 1 und (b) Bild 13; 3D-Rekonstruktionen: (c) Kameraperspektive für Bild 1 und (d) Kameraperspektive für Bild 13 (rot: Hintergrund, grün: IMO).

stellt, das auf einer geometrischen Segmentierung von Merkmalspunkten, die sowohl auf dem Hintergrund als auch auf unabhängig bewegten Objekten (IMOs) liegen, beruht. Als Klassifizierungsbasis wurden die Epipolargeometrien bzw. die algebraische Repräsentation in Form der Fundamentalmatrizen verwendet, die über einen iterativen RANSAC bestimmt wurden. Die Segmentierung wurde ausschließlich zwischen Schlüsselbildern, die über den GRIC-Ansatz detektiert worden sind, durchgeführt. In dem iterativen Prozess ist der Datensatz an Merkmalspunkten jeweils in korrekte Zuordnungen und Ausreißer unterteilt worden. Für die korrekten Zuordnungen wurde zusätzlich eine Abstandsmessung zum Schwerpunkt des segmentierten Datensatzes durchgeführt, um ggf. falsch klassifizierte Merkmalspunkte zu entfernen und somit die Robustheit des Verfahrens zu steigern. Ferner wurde guided-Matching zwischen Schlüsselbildern angewandt, um die Anzahl an Punktkorrespondenzen auf den IMOs zu erhöhen. Abschließend wurden die segmentierten Datensätze separat mittels SfM rekonstruiert. Das MSFM-Problem für dynamische Szenen wurde somit in einzelne SfM-Probleme zerlegt.

Diskussion: Die dargestellten Ergebnisse weisen zwar darauf hin, dass die geometrische Segmentierung und Rekonstruktion dynamischer Szenen prinzipiell über die Epipolargeometrie möglich ist. Allerdings stößt dieser theoretische Ansatz schnell an seine Grenzen, wenn bewegte Objekte, wie z. B. Personen, nicht mehr in sich selbst statisch sind. Zwar lassen sich auch diese in einzelne starre IMOs aufteilen, deren Größe würde aber voraussichtlich für die Rekonstruktion nicht mehr ausreichen, da sich nur noch wenige bzw. keine Merkmalspunkte auf diesen befinden. Für Merkmalspunkte, die überwiegend auf einer Ebene liegen, wie z. B. bei der *Box*-Sequenz, ist eine 3D-Rekonstruktion nicht möglich, da die Epipolargeometrie nicht eindeutig bestimmbar ist. Die Rekonstruktion des Hintergrundes und des Kamerapfads ist allerdings auch für dynamische Szenen möglich, wenn genügend Merkmalspunkte existieren und bewegte Objekte nicht zu große Bereiche der Bilder abdecken.

Ausblick: Eine vollständige Rekonstruktion von Szenen ist nur in wenigen Anwendungsbereichen erforderlich. Die Generierung von stereoskopischen Ansichten aus statischen Szenen mittels RSVS ist ein Beispiel dafür. Die Erweiterung von RSVS für dynamische Szenen könnte die Einschränkungen des Ansatzes reduzieren. Eine vollständige Segmentierung der bewegten Objekte wäre dafür sinnvoll, aber deren 3D-Rekonstruktion nicht zwingend notwendig. Wendet man weitere Segmentierungsverfahren auf die Bilder an, die nur diejenigen Objekte aus der Szene ausschneiden, bei denen Merkmalspunkte als IMO klassifiziert wurden, so könnten diese nachträglich interaktiv in virtuellen Stereobildern entsprechend ihrer Tiefe integriert werden. Das vorgestellte Verfahren kann somit als Vorstufe für Segmentierungsansätze dienen, um die 3D-Konvertierung von Videosequenzen weiter zu automatisieren.

9. Zusammenfassung und Ausblick

Das Ziel dieser Arbeit war die Entwicklung eines Konvertierungsverfahrens, welches aus monokularen Videosequenzen stereoskopische Sequenzen generiert. Das vorgestellte Konvertierungssystem basiert auf aktuellen Verfahren der Computer Vision und wurde um neue Ansätze erweitert und untersucht. Dabei wurden Bedingungen sowohl an die aufgenommenen Szenen als auch an die Kameraführung gestellt. Zum einen sollten die Szenen statisch sein und zum anderen musste sich die Kamera bei der Aufnahme bewegt haben, damit 3D-Informationen über binokulare Tiefenschätzverfahren zurückgewonnen werden konnten. Für die Synthese neuer virtueller Ansichten wurde ein Verfahren entwickelt, das unter Ausnutzung der ermittelten 3D-Informationen lediglich 2D-Transformationen zwischen benachbarten Bildern verwendet.

9.1. Zusammenfassung

In dieser Arbeit wurde eine umfangreiche Einführung in die geometrischen und algebraischen Grundlagen der Computer Vision gegeben, die für die Anforderungen einer 2D/3D-Konvertierung von essentieller Bedeutung sind. Einführend wurden die Eigenschaften der projektiven und der Epipolargeometrie beschrieben sowie Kameramodelle vorgestellt (Kapitel 2). Am Beispiel der Epipolargeometrie wurden robuste Schätzverfahren eingeführt, die eine große Bedeutung vor allem bei 3D-Rekonstruktionsansätzen besitzen. Diese Schätzverfahren wurden auf ihre Robustheit bezüglich Ausreißern und Rauscheinflüssen getestet und untersucht. Dabei hat sich RANSAC in den meisten Untersuchungsergebnissen als robustester Ansatz zur Eliminierung von Ausreißern erwiesen (Kapitel 3).

Weitere theoretische Grundlagen, die für die Rekonstruktion einer dreidimensionalen Szene und der Kameraparameter dienen, wurden vorerst für zwei Ansichten, die eine 3D-Szene aus unterschiedlichen Perspektiven abbilden, beschrieben (Kapitel 4), um eine fundierte Basis für die 3D-Rekonstruktion aus Videosequenzen zu schaffen. Dabei wurde zuerst auf die Korrespondenzanalyse eingegangen, mit der zusammenhängende Objekte zwischen den Ansichten gefunden werden können. Als Objekte wurden Merkmalspunkte gewählt, da diese speziell bei Videosequenzen einfach zugeordnet bzw. verfolgt werden können. Ferner wurden Verfahren vorgestellt, um die geometrischen Zusammenhänge

zwischen zwei Ansichten zu schätzen und die 3D-Koordinaten der Merkmalspunkte zu bestimmen.

In Kapitel 5 wurde die 3D-Rekonstruktion aus mehreren Ansichten mittels *Structure-from-Motion* (SfM) beschrieben. Für die Korrespondenzanalyse ist eine Erweiterung des LK-Trackers implementiert worden, der hinsichtlich der Auswirkungen verschiedener Kostenfunktionen und Bewegungsmodelle untersucht wurde. Für die Rekonstruktion sowohl der 3D-Struktur als auch der Kameraparameter wurde das *geometrisch robuste Informationskriterium* (GRIC) zur Detektion von Schlüsselbildern eingeführt und verschiedene lineare und nichtlineare Optimierungsansätze implementiert und verglichen.

Die Synthese virtueller stereoskopischer Ansichten wurde in Kapitel 6 anhand zweier verschiedener Ansätze vorgestellt. Der erste Ansatz gehört zu den bild- und tiefenbasierten Rendering-Methoden (DIBR), um Pixel entsprechend ihrer Tiefe in eine neue virtuelle Ansicht zu transformieren. Da für diesen Ansatz dichte Tiefenkarten erforderlich sind, wurde ein Überblick über die aktuellen Verfahren der Tiefenschätzung aus Bildern und Bildfolgen gegeben. Bei dem zweiten Verfahren (RSVS) wurde ein neuer Ansatz zur Erzeugung stereoskopischer Ansichten entwickelt, der ebenfalls zu den bildbasierten Rendering-Methoden zählt, allerdings keine dichten Tiefenkarten benötigt. Dieses Verfahren verwendet lediglich die Tiefeninformationen einzelner Merkmalspunkte, um globale Homographien zwischen virtuellen und Originalansichten zu berechnen. Die Homographien dienen als Transformationsvorschrift, so dass Pixel entsprechend der Transformationsparameter aus den Originalansichten in die virtuellen Ansichten transformiert werden können.

Dieser neuartige Ansatz wurde zusätzlich um eine Methode erweitert, mit der die Auflösung der Original- und der virtuellen Bilder durch Super-Resolution erhöht werden konnte. Dabei hat sich herausgestellt, dass sowohl Artefakte wie Bewegungsunschärfe und Spiegelungsfehler als auch Kodierungs- und Deinterlacing-Artefakte, die in Handkamerasequenzen häufig auftreten, reduziert werden konnten.

In Kapitel 7 wurden die Konvertierungsergebnisse von RSVS evaluiert. Für die Evaluierung sind zwei subjektive Qualitätstests nach den ITU-Empfehlungen durchgeführt worden, bei denen jeweils 15 Testpersonen den Qualitätseindruck der Testsequenzen mit Hilfe von Bewertungsskalen beurteilen sollten. In dem ersten Experiment wurde die Qualität von RSVS mittels der *Single-Stimulus* (SS) Methode untersucht. Insgesamt sind 25 Testsequenzen verglichen worden, die in drei verschiedenen Modi präsentiert wurden: 2D, 3D durch RSVS und 3D durch Zeitversatz. Bei dem zweiten Experiment wurden DIBR und RSVS mittels der *Double-Stimulus-Continous-Quality-Scale* (DSCQS) Methode untersucht. Als Referenz wurden drei Sequenzen von einem synthetischen 3D-Modell gerendert, so dass die Testpersonen jeweils eine Originalstereosequenz und eine konvertierte Stereosequenz anhand einer kontinuierlichen Bewertungsskala qualitativ beurteilen mussten.

Die Ergebnisse der beiden Tests belegen die hervorragenden Konvertierungseigenschaften von RSVS. Der erste Test, der vor allem allgemein die Qualität zwischen den Originalsequenzen und den konvertierten 3D-Sequenzen dokumentieren sollte, zeigte, dass kaum qualitative Einbußen durch die Konvertierung entstehen. Vielmehr wurden die konvertierten 3D-Sequenzen besser bewertet als die Originalsequenzen, was auf den räumlichen Wahrnehmungseffekt zurückzuführen ist. Bei dem zweiten Test zeigte sich deutlich, dass RSVS bessere Konvertierungsergebnisse liefert als DIBR, bei dem vor allem durch die Aufdeckungsproblematik die Qualität stark herabgesetzt wird.

In Kapitel 8 wurde eine Einführung in den Umgang mit dynamischen Szenen gegeben. Dabei wurde ein neues geometrisches Segmentierungsverfahren vorgestellt, welches Merkmalspunkte, die auf bewegten Szenenobjekten (IMOs) liegen, entsprechend detektiert und separat die 3D-Struktur ermittelt.

9.2. Wesentliche Beiträge dieser Arbeit

Die wesentlichen Beiträge und Ergebnisse dieser Arbeit sind hier noch einmal kurz zusammengefasst:

- ein umfangreicher Überblick über den Stand der Forschung
- eine umfangreiche Untersuchung verschiedener robuster Schätzverfahren
- eine genaue Untersuchung von Kostenfunktionen bei der Verfolgung von Merkmalspunkten
- ein erweiterter Ansatz zur Merkmalspunktverfolgung
- RSVS, ein neuer Ansatz zur Synthese stereoskopischer und Multiview-Ansichten
- eine Erweiterung von RSVS mittels Super-Resolution
- eine umfangreiche Evaluierung von RSVS mit standardkonformen subjektiven Qualitätstests
- ein neuer geometrischer Segmentierungsansatz zur 3D-Rekonstruktion dynamischer Szenen

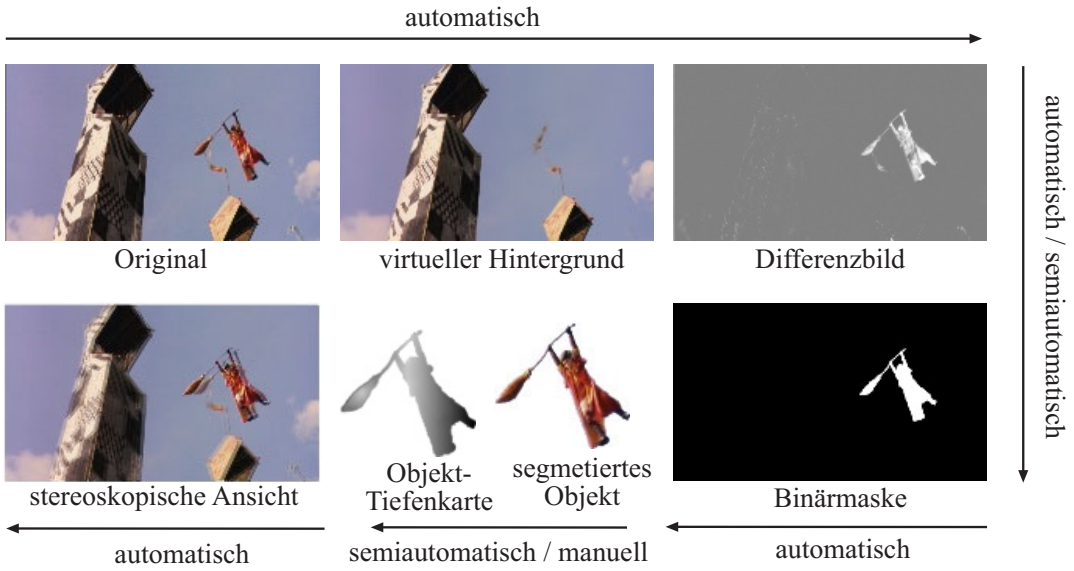


Abbildung 9.1.: Ausblick: Semiautomatische 2D/3D-Konvertierung mittels RSVS

9.3. Ausblick

Für den vorgestellten Ansatz zur Konvertierung von Film- oder Videosequenzen existieren eine Reihe von Erweiterungsmöglichkeiten, die hier kurz genannt werden sollen. Z. B. können bewegte Objekte durch Medianfilterung aus der Szene entfernt werden, indem - wie bei der Erstellung von Hintergrundmosaiken oder der Synthese einer virtuellen Ansicht bei RSVS - benachbarte Bilder einer Sequenz über Homographien auf die gleiche Ebene transformiert werden. Somit würde die statische Hintergrundszene mittels RSVS konvertiert werden. Die Objekte, die vorerst entfernt wurden, könnten nachträglich entsprechend ihrer Tiefe in die Stereosequenz integriert werden. Abbildung 9.1 zeigt den Ablauf einer möglichen 2D/3D-Konvertierung von dynamischen Szenen, die mit einer translatorisch bewegten Kamera aufgenommen wurden, am Beispiel einer Filmsequenz aus *Harry Potter*.

Zunächst wird das bewegte Objekt automatisch aus der Sequenz entfernt, so dass neben der Originalsequenz eine virtuelle Hintergrundsequenz existiert. Für den statischen Hintergrund lässt sich aufgrund der Kameraführung mittels RSVS eine stereoskopische Ansicht erstellen. Aus der Differenz zwischen dem Original und dem virtuellen Hintergrund lassen sich automatisch Objektsegmentierungsmasken generieren, die ggf. interaktiv nachbearbeitet werden müssen. Dem damit segmentierten Objekt wird in einem nächsten Schritt z. B. in einigen Schlüsselbildern manuell eine Tiefe zugewiesen, die in den Zwischenbildern interpoliert wird. Im letzten Schritt wird das Objekt der Tiefe entsprechend in die stereoskopische Hintergrundansicht gerendert.

Weitere Möglichkeiten, bewegte Objekte vorerst aus der Szene zu entfernen, lassen sich über Segmentierungsverfahren wie z. B. kontur- oder graphenbasierte Ansätze realisieren. Für vorwärts gerichtete Kamerabewegungen, bei denen RSVS zu große Transformationsfehler liefert, könnten Verfahren wie das *trifokale Warping* oder ggf. der Lumigraph zum Einsatz kommen.

Da ein Großteil der aufgenommenen Szenen nicht statisch ist bzw. die Kamera sich während der Aufnahme nicht immer translatorisch bewegt, müssen andere Ansätze gefunden werden, die die automatische Konvertierung ermöglichen bzw. den Konvertierungsprozess mit Hilfe semiautomatischer Verfahren zumindest beschleunigen. Hier werden hybride Ansätze immer mehr an Bedeutung gewinnen.

Bei feststehenden Kameras ist die Tiefenschätzung des Hintergrunds nur in einem Bild notwendig. Problematisch sind hier wieder bewegte Objekte in der Szene, die für jedes Bild der Sequenz segmentiert und entsprechend der Tiefe in die virtuelle Ansicht versetzt werden müssen. Da die Sgmentierungsmasken möglichst genau sein müssen, nimmt die Segmentierung den wohl bedeutendsten Anteil im Konvertierungsprozess ein.

Abschließend sei bemerkt, dass die 2D/3D-Konvertierung keine signifikanten Qualitätseinbußen gegenüber der ursprünglichen 2D-Version besitzen darf, da diese die Tiefenwahrnehmung stark beeinträchtigen und Nebenwirkungen wie Übelkeit oder Kopfschmerzen verursachen können, d.h. es sollten einzelne Szenen, die nur unter hohen Qualitätsverlusten konvertiert werden können, ggf. in 2D bleiben, da das Gehirn auch aus monokularen Informationen Tiefen wahrnehmen kann.

A. Anhang

A.1. Robuste Schätzung der Fundamentalmatrix

Tabelle A.1.: Vergleich robuster Schätzer bei verrauschten Datensätzen (Datensatz 1, Abstand zu Epipolarlinien): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400, Steigerung der Standardabweichung des gaußverteilten Rauschens von 0,25 bis 3, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler und der 2D-Reprojektionsfehler verwendet.

	Mittlere Abstand der Punkte zu den Epipolarlinien in Pixeln (±StdAbw. in Pixeln)						
StdAbw. in Pixeln	0,25	0,5	1,0	1,5	2,0	2,5	3,0
n8pt	**0,07** (0,02)	**0,24** (0,04)	0,58 (0,07)	0,79 (0,08)	0,88 (0,08)	0,94 (0,12)	**0,98** (0,11)
RANSAC-Samp	**0,07** (0,02)	**0,24** (0,05)	**0,54** (0,10)	**0,73** (0,15)	0,88 (0,21)	1,08 (0,26)	1,19 (0,27)
RANSAC-2Drep	**0,07** (0,02)	**0,24** (0,05)	0,55 (0,12)	0,74 (0,16)	**0,82** (0,19)	1,08 (0,21)	1,19 (0,30)
MLESAC-Samp	0,19 (0,06)	0,48 (0,33)	0,73 (0,22)	1,07 (0,39)	1,26 (0,33)	1,50 (0,41)	1,65 (0,41)
MLESAC-2Drep	0,20 (0,09)	0,44 (0,25)	0,81 (0,22)	0,97 (0,25)	1,31 (0,36)	1,45 (0,39)	1,54 (0,38)
MAPSAC-Samp	0,16 (0,05)	0,35 (0,11)	0,69 (0,23)	0,96 (0,32)	1,23 (0,35)	1,45 (0,36)	1,57 (0,38)
MAPSAC-2Drep	0,16 (0,05)	0,45 (0,31)	0,75 (0,25)	0,98 (0,28)	1,25 (0,32)	1,51 (0,47)	1,63 (0,39)
LMedS-Samp	0,08 (0,03)	0,23 (0,05)	0,58 (0,11)	0,78 (0,09)	0,87 (0,11)	**0,93** (0,13)	1,01 (0,13)
LMedS-2Drep	0,08 (0,02)	0,25 (0,06)	0,58 (0,10)	0,77 (0,09)	0,89 (0,11)	0,94 (0,14)	1,04 (0,17)

Tabelle A.2.: Vergleich robuster Schätzer bei verrauschten Datensätzen (Datensatz 2, Abstand zu Epipolarlinien): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 500, Steigerung der Standardabweichung des gaußverteilten Rauschens von 0,25 bis 3, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler verwendet.

	Mittlere Abstand der Punkte zu den Epipolarlinien in Pixeln (±StdAbw. in Pixeln)						
StdAbw. in Pixeln	0,25	0,5	1,0	1,5	2,0	2,5	3,0
n8pt	**0,04** (0,01)	**0,10** (0,03)	**0,24** (0,05)	**0,47** (0,11)	**0,66** (0,15)	**0,97** (0,29)	**1,10** (0,25)
RANSAC	**0,04** (0,01)	0,11 (0,04)	0,36 (0,15)	0,66 (0,24)	0,91 (0,25)	1,21 (0,35)	1,46 (0,50)
MLESAC	0,34 (0,49)	0,54 (0,39)	1,15 (1,09)	1,56 (1,43)	1,63 (0,74)	1,81 (0,51)	2,42 (0,77)
MAPSAC	0,19 (0,13)	0,63 (0,94)	1,20 (1,67)	1,53 (0,83)	1,65 (0,64)	1,94 (0,59)	2,08 (0,63)
LMedS	0,12 (0,39)	0,25 (0,37)	0,64 (0,89)	0,71 (0,34)	1,05 (0,65)	1,28 (0,68)	1,57 (1,41)

Tabelle A.3.: Vergleich robuster Schätzer bei verrauschten Datensätzen (Datensatz 1, Epipol-Abstand): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400, Steigerung der Standardabweichung des verteilten Rauschens von 0,25 bis 3, 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler und der 2D-Reprojektionsfehler verwendet.

	Mittlerer relativer Fehler zwischen dem korrekten Epipol und den berechneten Epipolen (±StdAbw.)						
StdAbw. in Pixeln	0,25	0,5	1,0	1,5	2,0	2,5	3,0
n8pt	**0,02** (0,01)	**0,08** (0,02)	0,39 (0,12)	0,59 (0,11)	0,63 (0,13)	0,71 (0,18)	0,72 (0,19)
RANSAC-Samp	**0,02** (0,01)	0,08 (0,03)	0,29 (0,11)	0,38 (0,20)	**0,31** (0,22)	**0,34** (0,26)	0,36 (0,28)
RANSAC-2Drep	**0,02** (0,01)	0,09 (0,03)	0,31 (0,15)	0,34 (0,21)	0,37 (0,29)	0,41 (0,29)	0,45 (0,31)
MLESAC-Samp	0,05 (0,04)	0,16 (0,22)	0,30 (0,27)	0,32 (0,23)	0,37 (0,27)	0,44 (0,30)	0,45 (0,31)
MLESAC-2Drep	0,04 (0,03)	0,13 (0,17)	0,31 (0,27)	0,32 (0,30)	0,51 (0,35)	0,44 (0,30)	**0,32** (0,25)
MAPSAC-Samp	0,03 (0,02)	0,08 (0,06)	0,26 (0,27)	**0,31** (0,26)	0,41 (0,30)	0,42 (0,30)	0,39 (0,28)
MAPSAC-2Drep	0,03 (0,02)	0,14 (0,16)	**0,26** (0,24)	0,32 (0,28)	0,34 (0,28)	0,34 (0,27)	0,50 (0,31)
LMedS-Samp	**0,02** (0,01)	0,08 (0,03)	0,35 (0,13)	0,55 (0,09)	0,62 (0,14)	0,67 (0,17)	0,69 (0,19)
LMedS-2Drep	**0,02** (0,01)	0,09 (0,04)	0,38 (0,14)	0,58 (0,12)	0,63 (0,15)	0,66 (0,17)	0,71 (0,19)

Tabelle A.4.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern (Datensatz 1: Abstand zu Epipolarlinien): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400, Steigerung des Ausreißeranteils von 5% bis 60%, $\sigma = 0,5$ Pixel (Standardabweichung des gaußverteilten Rauschens), 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler oder der 2D-Reprojektionsfehler verwendet.

	Mittlere Abstand der Punkte zu den Epipolarlinien in Pixeln (±StdAbw. in Pixeln)						
Ausreißer	20	40	80	120	160	200	240
n8pt	3,94 (0,00)	6,74 (0,00)	10,50 (0,00)	14,09 (0,00)	25,33 (0,00)	36,03 (0,00)	92,77 (0,00)
RANSAC-Samp	0,28 (0,17)	0,40 (0,16)	**0,36** (0,14)	0,48 (0,17)	0,39 (0,14)	0,49 (0,18)	**0,51** (0,20)
RANSAC-2Drep	0,25 (0,12)	**0,34** (0,12)	0,41 (0,15)	0,40 (0,16)	0,38 (0,17)	0,53 (0,21)	0,47 (0,18)
MLESAC-Samp	0,45 (0,22)	0,40 (0,18)	0,43 (0,16)	0,38 (0,11)	0,35 (0,11)	0,37 (0,14)	1,23 (0,69)
MLESAC-2Drep	0,47 (0,22)	0,45 (0,23)	0,45 (0,16)	**0,36** (0,11)	0,36 (0,10)	**0,33** (0,15)	1,23 (0,55)
MAPSAC-Samp	0,39 (0,19)	0,43 (0,19)	0,42 (0,19)	0,37 (0,13)	0,35 (0,13)	0,38 (0,14)	2,60 (3,47)
MAPSAC-2Drep	0,38 (0,13)	0,40 (0,22)	0,41 (0,18)	0,38 (0,16)	**0,29** (0,11)	0,44 (0,19)	2,40 (2,88)
LMedS-Samp	**0,24** (0,10)	0,39 (0,16)	0,57 (0,29)	2,67 (6,10)	9,26 (14,43)	27,89 (21,63)	68,15 (34,06)
LMedS-2Drep	0,27 (0,17)	0,40 (0,20)	0,44 (0,21)	0,76 (0,75)	0,81 (0,45)	7,13 (12,41)	40,92 (37,76)

Tabelle A.5.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern (Datensatz 2, Abstand zu Epipolarlinien): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 500, Steigerung des Ausreißeranteils von 5% bis 60%, $\sigma = 0,5$ Pixel (Standardabweichung des gaußverteilten Rauschens), 50 Durchläufe für jedes Verfahren.

	Mittlere Abstand der Punkte zu den Epipolarlinien in Pixeln (±StdAbw. in Pixeln)						
Ausreißer	25	50	100	150	200	250	300
RANSAC	**0,15** (0,07)	**0,22** (0,11)	**0,25** (0,21)	**0,35** (0,23)	**0,36** (0,17)	**0,34** (0,24)	**0,40** (0,20)
MLESAC	0.44 (0,23)	0.98 (1,71)	0.56 (0,45)	0.44 (0,24)	0.41 (0,25)	0.38 (0,35)	3.92 (1,06)
MAPSAC	0.60 (1,14)	0.86 (1,46)	0.52 (0,53)	0.39 (0,20)	0.38 (0,21)	0.48 (0,90)	5.15 (2,92)
LMedS	0.92 (2,68)	1.37 (2,23)	2.89 (11,13)	8.54 (23,62)	11.77 (25,68)	65.79 (56,11)	103.67 (46,70)

Tabelle A.6.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern (Datensatz 1: Epipol-Abstand): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400, Steigerung des Ausreißeranteils von 5% bis 60%, $\sigma = 0,5$ Pixel (Standardabweichung des gaußverteilten Rauschens), 50 Durchläufe für jedes Verfahren. Als Minimierungskriterium wurde der Sampson-Fehler oder der 2D-Reprojektionsfehler verwendet.

	Mittlerer relativer Fehler zwischen dem korrekten Epipol und den berechneten Epipolen (±StdAbw.)						
Ausreißer	20	40	80	120	160	200	240
n8pt	0,16 (0,00)	0,19 (0,00)	0,27 (0,00)	0,52 (0,00)	0,54 (0,00)	0,55 (0,00)	0,48 (0,00)
RANSAC-Samp	0,09 (0,09)	0,14 (0,12)	0,13 (0,05)	0,18 (0,10)	0,17 (0,10)	0,14 (0,07)	0,18 (0,08)
RANSAC-2Drep	0,08 (0,06)	**0,11** (0,04)	0,17 (0,12)	0,14 (0,07)	0,13 (0,09)	0,17 (0,11)	0,19 (0,14)
MLESAC-Samp	0,15 (0,18)	0,14 (0,17)	0,12 (0,11)	0,11 (0,09)	0,14 (0,09)	0,11 (0,06)	**0,14** (0,06)
MLESAC-2Drep	0,16 (0,16)	0,13 (0,12)	**0,12** (0,07)	**0,10** (0,06)	0,14 (0,12)	**0,09** (0,05)	0,16 (0,06)
MAPSAC-Samp	0,12 (0,12)	0,12 (0,11)	0,12 (0,10)	0,11 (0,08)	0,13 (0,10)	0,12 (0,07)	0,17 (0,08)
MAPSAC-2Drep	0,10 (0,08)	0,12 (0,13)	0,15 (0,15)	0,16 (0,16)	**0,10** (0,09)	0,13 (0,07)	0,17 (0,07)
LMedS-Samp	**0,07** (0,04)	0,14 (0,14)	0,33 (0,31)	0,37 (0,29)	0,41 (0,28)	0,41 (0,22)	0,39 (0,16)
LMedS-2Drep	0,10 (0,11)	0,19 (0,21)	0,18 (0,17)	0,27 (0,21)	0,38 (0,28)	0,49 (0,30)	0,41 (0,27)

Tabelle A.7.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern (Datensatz 1, Abstand zu Epipolarlinien): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 400, Steigerung des Ausreißeranteils von 5% bis 60%, $\sigma = 2,0$ Pixel (Standardabweichung des gaußverteilten Rauschens), 50 Durchläufe für jedes Verfahren.

	Mittlere Abstand der Punkte zu den Epipolarlinien in Pixeln (±StdAbw. in Pixeln)						
Ausreißer	25	50	100	150	200	250	300
RANSAC	0,97 (0,18)	0,88 (0,15)	**0,83** (0,18)	**0,92** (0,27)	**1,02** (0,26)	1,02 (0,20)	**1,01** (0,27)
MLESAC	1,23 (0,25)	1,19 (0,28)	1,08 (0,24)	1,10 (0,30)	1,04 (0,23)	**0,95** (0,25)	1,05 (0,38)
MAPSAC	1,19 (0,31)	1,15 (0,27)	1,08 (0,22)	1,01 (0,26)	1,22 (1,15)	2,77 (1,77)	2,25 (3,45)
LMedS	**0,88** (0,13)	**0,85** (0,18)	2,29 (5,86)	7,53 (13,94)	7,01 (9,59)	25,71 (22,34)	60,18 (30,66)

Tabelle A.8.: Vergleich robuster Schätzer bei der Eliminierung von Ausreißern (Datensatz 2, Abstand zu Epipolarlinien): Gesamtanzahl an Punktkorrespondenzen $\widehat{=}$ 500, Steigerung des Ausreißeranteils von 5% bis 60%, $\sigma = 2,0$ Pixel (Standardabweichung des gaußverteilten Rauschens), 50 Durchläufe für jedes Verfahren.

	Mittlere Abstand der Punkte zu den Epipolarlinien in Pixeln (±StdAbw. in Pixeln)						
Ausreißer	25	50	100	150	200	250	300
RANSAC	**1,01** (0,29)	**1,18** (0,38)	**0,81** (0,24)	**1,11** (0,28)	**1,15** (0,30)	**0,92** (0,34)	**1,04** (0,33)
MLESAC	1,82 (0,89)	1,78 (0,50)	1,42 (0,49)	1,49 (0,58)	1,44 (0,74)	1,40 (0,86)	3,69 (0,66)
MAPSAC	1,58 (0,51)	1,80 (0,54)	1,58 (0,47)	1,63 (0,59)	2,48 (2,59)	3,27 (2,08)	5,67 (3,32)
LMedS	1,50 (0,95)	2,42 (3,04)	2,32 (1,34)	16,82 (32,63)	24,97 (37,14)	51,18 (51,65)	97,45 (38,31)

A.2. Levenberg-Marquardt-Algorithmus

Im folgendem sind die Berechnungsschritte des LM-Algorithmus zusammengefasst:

Ziel: Bestimmung des optimalen Parametervektors $\mathbf{p}^{+} \in \mathbb{R}^m$, der den quadratischen Fehler $|\mathbf{x} - f(\mathbf{p})|^2$ minimiert.

Gegeben: Eine Vektorfunktion $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$, $n \geq m$, ein initialer Parametervektor $\mathbf{p}_0 \in \mathbb{R}^m$ sowie ein Vektor mit Messwerten $\mathbf{x} \in \mathbb{R}^n$.

$k := 0$, $v := 2$, $\mathbf{p} := \mathbf{p}_0$, $A := J^T J$, $\epsilon_\mathbf{p} := x - f(\mathbf{p})$, $\mathbf{g} := J^T \epsilon$, $\mu = \tau \; max_{i=1,\ldots,m}(A_{ii})$
Wiederhole solange $(|\mathbf{g}|_\infty > \varepsilon_1)$ und $(k < k_{max})$
    $k := k + 1$
    Wiederhole
        Bestimme $\delta_\mathbf{p}$ aus $(A + \mu \underline{I})\delta_\mathbf{p} = \mathbf{g}$
        Wenn $(|\delta_\mathbf{p}| \leq \varepsilon_2 \; |\mathbf{p}|)$ *break* := true
        sonst
            $\mathbf{p}_{neu} := \mathbf{p} + \delta_\mathbf{p}$
            $\rho := (|\epsilon_\mathbf{p}|^2 - |\mathbf{x} - f(\mathbf{p}_{neu})|^2)/(\delta_\mathbf{p}^T(\mu \; \delta_\mathbf{p} + \mathbf{g}))$
            Wenn $(\rho > 0)$
                $\mathbf{p} := \mathbf{p}_{neu}$, $A := J^T J$, $\epsilon_\mathbf{p} := \mathbf{x} - f(\mathbf{p})$, $\mathbf{g} := J^T \epsilon_\mathbf{p}$
                break := $(|\mathbf{g}|_\infty \leq \varepsilon_1)$,
                $\mu := \mu \; \max(\frac{1}{3}, 1 - (2\,\rho - 1)^3))$, $v := 2$
            sonst $\mu := \mu v$, $v := 2v$
    bis ($\rho > 0$ oder *break*)
Ende der Schleife
$\mathbf{p}^{+} := \mathbf{p}$

Ergebnis: optimierter Parametervektor $\mathbf{p}^{+}$.

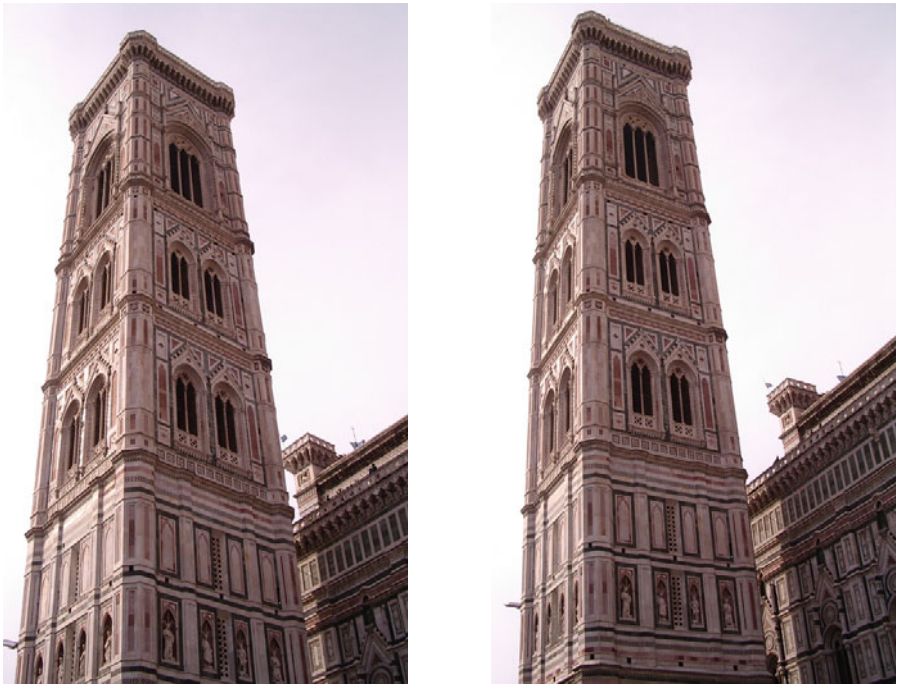
Algorithmus 12: LM-Algorithmus (Quelle: [115])

A.3. Verwendete Bilder und Sequenzen

A.3.1. Verwendete Bilder in Kapitel 3 und 4

(a) (b)

Abbildung A.1.: Bildpaar *Halle*



(a) (b)

Abbildung A.2.: Bildpaar *Turm*

(a) (b)

Abbildung A.3.: Bildpaar *Korridor*

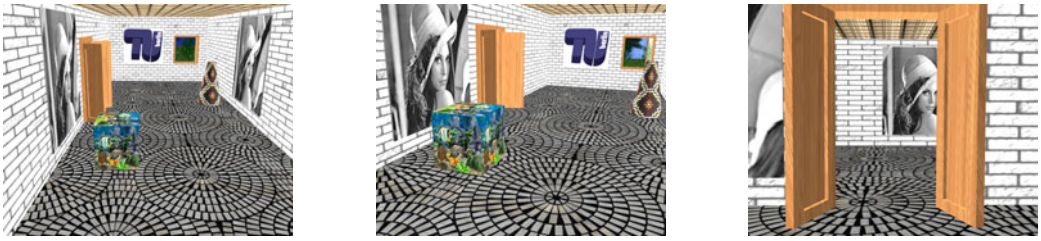
(a) (b)

Abbildung A.4.: Bildpaar *Rathaus*

(a) (b)

Abbildung A.5.: Bildpaar *Totem*

A.3.2. Verwendete Videosequenzen in Kapitel 5, 6 und 7



(a) Bild 1 (b) Bild 160 (c) Bild 320

Abbildung A.6.: Beispielbilder der Sequenz *TUBroom* (Auflösung: 720 × 576, 320 Bilder)

(a) Bild 1 (b) Bild 100 (c) Bild 208

Abbildung A.7.: Beispielbilder der Sequenz *Palast* (Auflösung: 720 × 576, 208 Bilder)

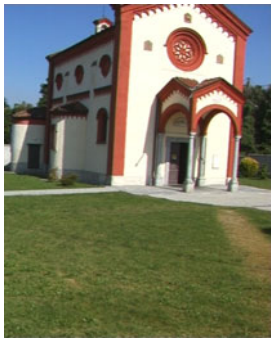
(a) Bild 1 (b) Bild 60 (c) Bild 109

Abbildung A.8.: Beispielbilder der Sequenz *Felsen* (Auflösung: 720 × 576, 109 Bilder)

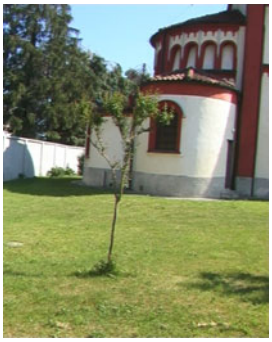
(a) Bild 1 (b) Bild 700 (c) Bild 1400

Abbildung A.9.: Beispielbilder der Sequenz *Dom* (Auflösung: 720 × 576, 1400 Bilder)

(a) Bild 1



(b) Bild 200



(c) Bild 555

Abbildung A.10.: Beispielbilder der Sequenz *Kirche* (Auflösung: 576 × 720, 555 Bilder)

(a) Bild 1

(b) Bild 92

(c) Bild 186

Abbildung A.11.: Beispielbilder der Sequenz *Medusa* (Auflösung: 448 × 358, 186 Bilder)

(a) Bild 1

(b) Bild 700

(c) Bild 1012

Abbildung A.12.: Beispielbilder der Sequenz *Fassade* (Auflösung: 720 × 576, 1012 Bilder)

(a) Bild 1

(b) Bild 315

(c) Bild 630

Abbildung A.13.: Beispielbilder der Sequenz *Statue* (Auflösung: 720 × 576, 630 Bilder)

BBC-Dokumentation *Planet Erde*

(a) Bild 1 (b) Bild 200 (c) Bild 408

Abbildung A.14.: Beispielbilder der Sequenz *Höhle* aus der Serie *Höhlenwelten* (Auflösung: 720 × 404, 408 Bilder)

(a) Bild 1 (b) Bild 100 (c) Bild 200

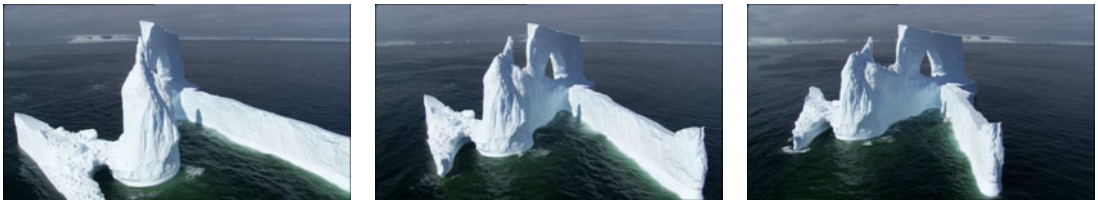
Abbildung A.15.: Beispielbilder der Sequenz *Monument Valley* aus der Serie *Wüstenwelten* (Auflösung: 720 × 404, 200 Bilder)

(a) Bild 1 (b) Bild 150 (c) Bild 262

Abbildung A.16.: Beispielbilder der Sequenz *Wald* aus der Serie *Waldwelten* (Auflösung: 720 × 404, 262 Bilder)

(a) Bild 1 (b) Bild 80 (c) Bild 159

Abbildung A.17.: Beispielbilder der Sequenz *Riff* aus der Serie *Wasserwelten* (Auflösung: 720 × 404, 159 Bilder)



(a) Bild 1 (b) Bild 140 (c) Bild 282

Abbildung A.18.: Beispielbilder der Sequenz *Eisberg* aus der Serie *Eiswelten* (Auflösung: 720 × 404, 282 Bilder)

(a) Bild 1 (b) Bild 200 (c) Bild 330

Abbildung A.19.: Beispielbilder der Sequenz *Dschungel* aus der Serie *Urwaldwelten* (Auflösung: 720 × 404, 330 Bilder)

(a) Bild 1 (b) Bild 115 (c) Bild 230

Abbildung A.20.: Beispielbilder der Sequenz *Berge* aus der Serie *Bergwelten* (Auflösung: 720 × 404, 230 Bilder)

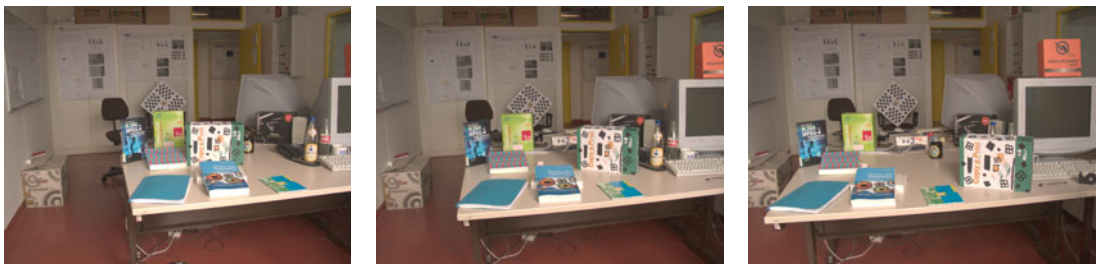
(a) Bild 1 (b) Bild 80 (c) Bild 140

Abbildung A.21.: Beispielbilder der Sequenz *Tiefsee* aus der Serie *Tiefseewelten* (Auflösung: 720 × 404, 140 Bilder)

A.3.3. Verwendete Sequenzen in Kapitel 8

(a) Bild 1 (b) Bild 120 (c) Bild 240

Abbildung A.22.: Beispielbilder der Sequenz *TUBroom-IMO* (Auflösung: 720 × 576, 240 Bilder)



(a) Bild 1 (b) Bild 5 (c) Bild 13

Abbildung A.23.: Beispielbilder der Stop-Motion-Sequenz *Box* (Auflösung: 720 × 535, 13 Bilder)

(a) Bild 1 (b) Bild 6 (c) Bild 13

Abbildung A.24.: Beispielbilder der Stop-Motion-Sequenz *Sofa* (Auflösung: 1024 × 768, 13 Bilder)

A.4. Informationsbögen für die subjektiven Qualitätstests

Subjektive Qualitätsmessung

Ihnen werden 25 Sequenzen gezeigt. Diese Sequenzen werden jeweils in untereinander wechselnder Reihenfolge dargestellt:

- 2D
- 3D Methode 1
- 3D Methode 2

Sie sollen jede dieser Sequenzen subjektiv nach ihrer Qualität bewerten. Dabei steht vor allem der Gesamteindruck der Sequenz im Vordergrund.

Hinweis: Der Versuchsaufbau beinhaltet u.a. eine Shutter-Brille. Bitte halten Sie keine Gegenstände zwischen die Brille und den Sender, da es sonst zu Synchronisierungsproblemen kommen kann.

Der Versuchsablauf sieht zwei Phasen vor: eine kurze Trainingsphase zur Vorbereitung und eine Evaluierungsphase von etwa 27 Minuten. Sowohl in der Trainingsphase als auch in der Evaluierungsphase stehen 10 Sekunden Zeit für die Bewertung der voran gezeigten Sequenz zur Verfügung. Dazu soll ein Kreuz auf der Bewertungsskala (s.u.) bei der entsprechenden Zahl gemacht werden. Nach den 10 Sekunden folgt die nächste Sequenz usw.

Bewertungsskala

	11
Excellent	10
	9
Good	8
	7
Fair	6
	5
Poor	4
	3
Bad	2
	1

Abbildung A.25.: Informationsbogen Qualitätstest I

Subjektive Qualitätsmessung

Ihnen werden mehrere synthetische 3D Sequenzen gezeigt. Jede dieser Sequenzen wird zweimal im Wechsel mit einer Referenzsequenz gezeigt (welches die Referenzsequenz ist, wird nicht angegeben).

Sie sollen sowohl jeweils die Referenzsequenz als auch die Testsequenz subjektiv nach ihrer Qualität bewerten.

Der Versuchsablauf sieht zwei Phasen vor: eine kurze Trainingsphase zur Vorbereitung und eine Evaluierungsphase von etwa 12 Minuten. Sowohl in der Trainingsphase als auch in der Evaluierungsphase stehen 10 Sekunden Zeit für die Bewertung der voran gezeigten Sequenzen zur Verfügung. Dazu soll ein waagerechter **Strich** auf der Bewertungsskala (s.u.) an der entsprechenden Stelle gemacht werden. Nach den 10 Sekunden folgt die nächste Sequenz usw.

Ablauf: Sequenz A → 2 s → Sequenz B → 2 s → Sequenz A → 2 s → Sequenz B → 10 s

Bewertungsskala

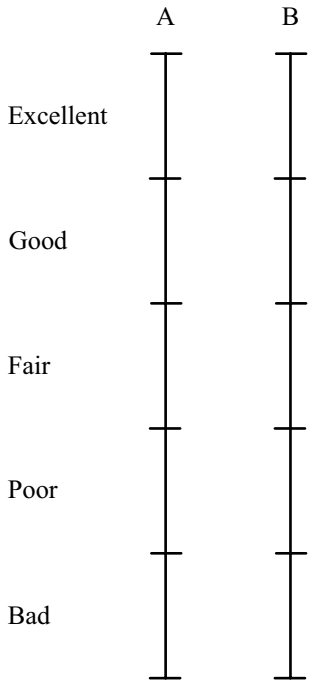


Abbildung A.26.: Informationsbogen Qualitätstest II

Literaturverzeichnis

[1] Adelson, E. ; Bergen, J. : The Plenoptic Function and the Elements of Early Vision. In: M.Landy (Hrsg.) ; Movshon, J. (Hrsg.): *Computation Models of Visual Processing*. MIT Press, Cambridge, 1991, S. 3–20

[2] Akbarzadeh, A. ; Frahm, J.-M. ; Mordohai, P. ; Clipp, B. ; Engels, C. ; Gallup, D. ; Merrell, P. ; Phelps, M. ; Sinha, S. ; Talton, B. ; Wang, L. ; Yang, Q. ; Stewenius, H. ; Yang, R. ; Welch, G. ; Towles, H. ; Nistér, D. ; Pollefeys, M. : Towards Urban 3D Reconstruction From Video. In: *3rd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*. Chapel Hill, USA : IEEE Computer Society, 2006

[3] Arce, E. ; Marroquin, J. : High-precision Stereo Disparity Estimation Using HMMF Models. In: *Image and Vision Computing* 25 (2007), May, Nr. 5, S. 623–636

[4] Asada, H. ; Brady, M. : The Curvature Primal Sketch. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8 (1986), Nr. 1, S. 2–14

[5] Atzpadin, N. ; Kauff, P. ; Schreer, O. : Stereo Analysis by Hybrid Recursive Matching for Real-Time Immersive Video Conferencing. In: *IEEE Transactions on Circuits and Systems for Video Technology* 14 (2004), March, Nr. 3, S. 321–333

[6] Ayache, N. ; Hansen, C. : Rectification of Images for Binocular and Trinocular Stereovision. In: *Proc. of International Conference on Pattern Recognition (ICPR)*, 1988, S. 11–16

[7] Baker, S. ; Nayar, S. ; Murase, H. : Parametric Feature Detection. In: *International Journal of Computer Vision* 27 (1998), Nr. 1, S. 27–50

[8] Battiato, S. ; Capra, A. ; S.Curti ; Cascia, M. : 3D Stereoscopic Image Pairs by Depth-Map Generation. In: *2nd International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT)*, 2004

[9] Beardsley, P. ; Zisserman, A. ; Murray, D. : Sequential Updating of Projective and Affine Structure from Motion. In: *International Journal of Computer Vision* 23 (1997), Nr. 3, S. 235–259

[10] Beardsley, P. A. ; Zisserman, A. ; Murray, D. W.: Navigation Using Affine Structure from Motion. In: *Proc. of European Conference on Computer Vision (ECCV), LNCS Series* Bd. 801, 1994, S. 85–96

[11] Beaudet, P. R.: Rotational Invariant Image Operators. In: *Proc. of International Conference on Pattern Recognition (ICPR)*, 1978, S. 579–583

[12] Benzie, P. ; Watson, J. ; Surman, P. ; Rakkolainen, I. ; Hopf, K. ; Urey, H. ; Sainov, V. ; von Kopylow, C. : A Survey of 3DTV Displays: Techniques and Technologies. In: *IEEE Transactions on Circuits and Systems for Video Technology* 17 (Nov. 2007), Nr. 11, S. 1647–1658

[13] Beraldin, J.-A. ; Picard, M. ; El-Hakim, S. ; Godin, G. ; Borgeat, L. ; Blais, F. ; Paquet, E. ; Rioux, M. ; Valzano, V. ; Bandiera, A. : Virtual Reconstruction of Heritage Sites: Opportunities and Challenges Created by 3D Technologies. In: *International Workshop on Recording, Modeling and Visualization of Cultural Heritage*. Ascona, Switzerland, May 2005

[14] Blinn, J. F. ; Newell, M. E.: Texture and Reflection in Computer Generated Images. In: *Communications of the ACM (CACM)* 19 (1976), October, Nr. 10, S. 542–547

[15] van den Boomgaard, R. ; van de Weijer, J. : Robust Estimation of Orientation for Texture Analysis. In: *Texture 2002, 2nd International Workshop on Texture Analysis and Synthesis*. Copenhagen, June 2002, S. 135–138

[16] Borman, S. ; Stevenson, R. L.: Super-Resolution from Image Sequences - A Review. In: *Midwest Symposium on Systems and Circuits*, 1998, S. 374–378

[17] Bouguet, J.-Y. : Pyramidal Implementation of the Lucas Kanade Feature Tracker. Description of the Algorithm. / Intel Corporation, Microprocessor Research Labs. 2000. – Forschungsbericht

[18] Börner, R. ; Duckstein, B. ; Machui, O. ; Röder, H. ; Sinnig, T. ; Sikora, T. : A Family of Single-User Autostereoscopic Displays with Head-Tracking Capabilities. In: *IEEE Transactions on Circuits and Systems for Video Technology* 10 (2000), Nr. 2, S. 234–243

[19] Broggi, A. ; Caraffi, C. ; Fedriga, R. I. ; Grisleri, P. : Obstacle Detection with Stereo Vision for off-road Vehicle Navigation. In: *IEEE International Workshop on Machine Vision for Intelligent Vehicles*. San Diego, USA, June 2005

[20] Buehler, C. ; Matusik, W. ; McMillan, L. : Creating and Rendering Image-Based Visual Hulls. 1999 (MIT, TR-780). – Forschungsbericht. – 18 S

[21] Carranza, J. ; Theobalt, C. ; Magnor, M. A. ; Seidel, H.-P. : Free-Viewpoint Video of Human Actors. In: *ACM Transactions on Graphics (Proc. of ACM Siggraph)* Bd. 22. San Diego, USA, July 2003, S. 569–577

[22] Chen, W.-Y. ; Chang, Y.-L. ; Lin, S.-F. ; Ding, L.-F. ; Chen, L.-G. : Efficient Depth Image Based Rendering with Edge Dependent Depth Filter and Interpolation. In: *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, 2005, S. 1314–1317

[23] Chum, O. ; Matas, J. : Matching with PROSAC - Progressive Sample Consensus. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005

[24] Cornelis, N. ; Leibe, B. ; Cornelis, K. ; Gool, L. J. V.: 3D City Modeling Using Cognitive Loops. In: *3rd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*. Chapel Hill, USA : IEEE Computer Society, 2006, S. 9–16

[25] Costeira, J. ; Kanade, T. : A Multibody Factorization Method for Independently Moving Objects. In: *International Journal of Computer Vision (IJCV), special issue on Dynamic Vision* 29 (1998), September, Nr. 3, S. 159–179

[26] Csurka, G. ; Demirdjian, D. ; Horaud, R. : Finding the Collineation Between Two Projective Reconstructions. In: *Computer Vision and Image Understanding* 75 (1999), Nr. 3, S. 260–268

[27] Curti, S. ; Sirtori, D. ; Vella, F. : 3D Effect Generation from Monocular View. In: *1st International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT)*, 2004

[28] Deriche, R. ; Blaszka, T. : Recovering and Characterizing Image Features Using an Efficient Model Based Approach. In: *International Conference on Patern Recognition (ICPR)*, 1993, S. 530–535

[29] Deriche, R. ; Giraudon, G. : 2-D Curve Matching Using High Curvature Points: Application to Stereo Vision. In: *International Conference on Patern Recognition (ICPR)*, 1990, S. 240–242

[30] Deriche, R. ; Giraudon, G. : A Computational Approach for Corner and Vertex Detection. In: *International Journal of Computer Vision* 10 (1993), Nr. 2, S. 101–124

[31] Diplaris, S. ; Grammalidis, N. ; Tzovaras, D. ; Strintzis, M. G.: Generation of Stereoscopic Image Sequences Using Structure and Rigid Motion Estimation by Extended Kalman Filters. In: *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*. Lausanne, Switzerland, 2002

[32] Dreschler, L. ; Nagel, H. H.: Volumetric Model and 3D Trajectory of a Moving Car Derived from Monocular TV Frame Sequences of a Street Scene. In: *Computer Graphics and Image Processing* 20 (1982), S. 199–228

[33] Durou, J.-D. ; Falcone, M. ; Sagona, M. : Numerical Methods for Shape-from-Shading: A New Survey with Benchmarks. In: *Computer Vision and Image Understanding* 109 (2008), January, Nr. 1, S. 22–43

[34] Ebers, O. : *Überblick über aktuelle Verfahren zur Tiefenschätzung aus 2D-Video-Sequenzen*, Technische Universität Berlin, Studienarbeit, March 2004

[35] Falkenhagen, L. : Depth Estimation from Stereoscopic Image Pairs Assuming Piecewise Continuous Surfaces. In: *European Workshop on combined Real and Synthetic Image Processing for Broadcast and Video Production*. Hamburg, November 1994

[36] Falkenhagen, L. : Block-Based Depth Estimation from Image Triples with Unrestricted Camera Setup. In: *IEEE Workshop on Multimedia Signal Processing*. Princeton, New Jersey, USA, June 1997

[37] Falkenhagen, L. : Hierarchical Block-based Disparity Estimation Considering Neighbourhood Constraints. In: *International workshop on SNHC and 3D Imaging*, 1997

[38] Faugeras, O. : *Three Dimensional Computer Vision: A Geometric Viewpoint*. MIT-Press, Cambridge, 1993

[39] Faugeras, O. ; Luong, Q.-T. : *The Geometry of Multiple Images*. Massachusetts Institute of Technology, 2001

[40] Faugeras, O. D. ; Luong, Q.-T. ; Maybank, S. J.: Camera Self-Calibration: Theory and Experiments. In: *Proc. of European Conference on Computer Vision (ECCV)*. London, UK : Springer-Verlag, 1992, S. 321–334

[41] Favaro, P. : Shape from Focus and Defocus: Convexity, Quasiconvexity and Defocus-Invariant Textures. In: *IEEE International Conference on Computer Vision (ICCV)* (14-21 Oct. 2007), S. 1–7

[42] Fehn, C. ; Kauff, P. ; de Beeck, M. O. ; Ernst, F. ; Ijsselsteijn, W. ; M. Pollefeys, L. V. ; Ofek, E. ; Sexton, I. : An Evolutionary and Optimised Approach on 3D-TV. In: *Proc. of IBC 2002, International Broadcast Convention*. Amsterdam, Netherlands, September 2002

[43] Fehn, C. : A 3D-TV System Based on Video Plus Depth Information. In: *Asilomar Conference on Signals, Systems and Computer*. Pacific Grove, CA, USA, November 2003

[44] Fehn, C. : Depth-Image-Based Rendering, Compression and Transmission for a New Approach on 3D-TV. In: *Proc. of SPIE: Stereoscopic Displays and Applications*. San Jose, CA, USA, January 2004

[45] Fehn, C. : *Depth-Image-Based Rendering (DIBR), Compression, and Transmission for a Flexible Approach on 3DTV*, Technische Universität Berlin, Diss., 2006

[46] Feldmann, I. ; Schreer, O. ; Kauff, P. : Nonlinear Depth Scaling For Immersive Video Applications. In: *Proc. of International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. London, UK, April 2003, S. 433–438

[47] Feng, C. ; Hung, Y. : A Robust Method for Estimating the Fundamental Matrix. In: *VIIth Digital Image Computing: Techniques and Applications*, 2003

[48] Fischler, M. ; Bolles, R. : RANdom SAmpling Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography. In: *Commun. Assoc. Comp. Mach.* 24 (1981), S. 381–395

[49] Fitzgibbon, W. ; Zisserman, A. : Multibody Structure and Motion: 3D Reconstruction of Independently Moving Objects. In: *Proc. of European Conference on Computer Vision (ECCV)*. Dublin, Ireland, June 26 - July 1

[50] Förstner, W. : A Framework for Low Level Feature Extraction. In: *Proc. of European Conference on Computer Vision (ECCV)*, 1994, S. 383–394

[51] Fusiello, A. ; Roberto, V. ; Trucco, E. : Efficient Stereo with Multiple Windowing. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Puerto Rico, 1997, S. 858–863

[52] Fusiello, A. ; Roberto, V. ; Trucco, E. : Symmetric Stereo with Multiple Windowing. In: *International Journal of Pattern Recognition and Artificial Intelligence* 14 (2000), S. 1053–1066

[53] Fusiello, A. ; Trucco, E. ; Roberto, V. : Improving Feature Tracking with Robust Statistics. In: *Pattern Analysis and Application* 2 (1999), S. 312–320

[54] G. Qian, Q. Z.: Automated Reconstruction of 3D Models from Real Environments. In: *IEEE Transactions on Image Processing* 14 (2005), January, Nr. 1

[55] Garcia, B. J.: Approaches to Stereoscopic Video Based on Spatio-Temporal Interpolation. In: *Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems IV*. San Jose, CA, USA, 1997

[56] Godin, G. ; Beraldin, J.-A. ; Taylor, J. ; Cournoyer, L. ; Rioux, M. ; El-Hakim, S. ; Baribeau, R. ; Blais, F. ; Boulanger, P. ; Picard, M. ; J., D. : Active Optical 3D Imaging for Heritage Applications. In: *IEEE Computer Graphics and Applications* 22 (2002), Nr. 5, S. 24–36

[57] Golub, G. ; Loan, C. V.: *Matrix Computations*. 3. Auflage. John Hopkins University Press, 1996

[58] Gonçalves, J. ; Sequeira, V. : Sensor-Based Depth Capturing. In: Schreer, O. (Hrsg.) ; Kauf, P. (Hrsg.) ; Sikora, T. (Hrsg.): *3D Videocommunication - Algorithms, concepts and real-time systems in human centered communication*. John Wiley & Sons, Ltd, 2005, S. 299–313

[59] Gortler, S. J. ; Grzeszczuk, R. ; Szeliski, R. ; Cohen, M. F.: The Lumigraph. In: *SIGGRAPH: Proc. of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA : ACM, 1996, S. 43–54

[60] Grau, O. ; Koch, R. ; Lavagetto, F. ; Sarti, A. ; Tubaro, S. ; Woetzel, J. : The ORIGAMI Project: Advanced Tools for Creating and Mixing Real and Virtual Content in Film and TV Production. In: *IEE Proceedings: Vision, Image and Signal Processing* 152 (2005), August, Nr. 4, S. 454–469

[61] Hager, G. D. ; Belhumeur, P. N.: Efficient Region Tracking with Parametric Models of Geometry and Illumination. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (1998), Nr. 10, S. 1025–1039

[62] Han, M. ; Kanade, T. : Perspective Factorization Methods for Euclidean Reconstruction / Robotics Institute, Carnegie Mellon University. Pittsburgh, PA, August 1999 (CMU-RI-TR-99-22). – Forschungsbericht

[63] Harman, P. ; Flack, J. ; Fox, S. ; Dowley, M. : Rapid 2D to 3D Conversion. In: *Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems IX* Bd. 4660. San José, CA, USA, 2002, S. 78–86

[64] Harris, C. ; Stephens, M. : A Combined Corner and Edge Detector. In: *4th Alvey Vision Conference*, 1988, S. 147–151

[65] Hartley, M. : Estimation of Relative Camera Positions for Uncalibrated Cameras. In: in Computer Science 588, L. N. (Hrsg.): *Proc. of Eeuropean Conference on Computer Vision (ECCV)*, 1992, S. 579–587

[66] Hartley, R. : In Defense of the Eight-Point Algorithm. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 19 (1997), October, Nr. 6, S. 580–593

[67] Hartley, R. ; Sturm, P. : Triangulation. In: *Computer Vision and Image Understanding* 68 (1997), November, Nr. 2, S. 146–157

[68] Hartley, R. ; Zisserman, A. : *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003

[69] Hetzheim, H. ; Börner, A. : Vehicle Detection from Airborne Images by Separation of Texture Properties and their Fusion. In: *Image and Vision Computing Conference*. New Zealand, 2003

[70] Hoiem, D. ; Efros, A. A. ; Hebert, M. : Automatic Photo Pop-up. In: *ACM Transactions on Graphics* 24 (2005), Nr. 3, S. 577–584

[71] Hu, X. ; Ahuja, N. : Matching Point Features with Ordered Geometric, Rigidity and Disparity Constraints. In: *Trans. on Pattern Analysis and Machine Intelligence* 16 (1994), October, Nr. 10

[72] Huber, P. J.: *Robust Statistics*. John Wiley & Sons, 1981

[73] Hucka, M. ; Kaplan, S. : Texture-Based Processing in Early Vision and a Proposed Role for Coarse-Scale Segmentation. In: *Proc. of Midwest Artificial Intelligence and Cognitive Science Conference (MAICS)*. Bloomington, Indiana, USA, 1996

[74] Iddan, G. J. ; Yahav, G. : 3D Imaging in the Studio. In: *Proc. of SPIE Videometrics and Optical Methods for 3D Shape Measurements*. San Jose, CA, USA, January 2001, S. 48–55

[75] IJsselsteijn, W. ; de Ridder, H. ; Vliegen, J. : Effects of Stereoscopic Filming Parameters and Exposure Duration on the Subjective Assessment of Eye Strain. In: *Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems VII* Bd. 3957. San Jose, CA, USA, April 2000, S. 12–22

[76] Imre, E. ; Knorr, S. ; Alatan, A. A. ; Sikora, T. : Prioritized Sequential 3D Reconstruction in Video Sequences of Dynamic Scenes. In: *IEEE Conference on Image Processing (ICIP)*. Atlanta, GA, USA, October 2006

[77] Imre, E. ; Knorr, S. ; Özkalayci, B. ; Topay, U. ; Alatan, A. A. ; Sikora, T. : Towards 3-D scene reconstruction from broadcast video. In: *Signal Processing: Image Communication* 22 (2007), S. 108–126

[78] Irani, M. ; Anandan, P. : A Unified Approach to Moving Object Detection in 2D and 3D Scenes. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (1998), June, Nr. 6, S. 577–589

[79] Irani, M. ; Peleg, S. : Super Resolution from Image Sequences. In: *Proc. of International Conference on Pattern Recognition (ICPR)*. Atlantic City, NJ, 1990

[80] ISO/IEC, .-. . *Information Technology – Coding of Audio-Visual Objects – Part 16: Animation Framework eXtension (AFX)*. 2006

[81] ITU: *Methology for the Subjective Assessment of the Quality of Television Pictures*. 2000. – ITU-R BT.500-10

[82] ITU: *Subjective Assessment of Stereoscopic Television Pictures*. 2000. – ITU-R BT.1438

[83] Ivekovic, S. ; Fusiello, A. ; Trucco, E. : Fundamentals of Multiple-view Geometry. In: Schreer, O. (Hrsg.) ; Kauf, P. (Hrsg.) ; T, S. (Hrsg.): *3D Videocommunication - Algorithms, concepts and real-time systems in human centered communication*. John Wiley & Sons, Ltd, 2005, S. 93–114

[84] Jaureguizar, F. ; Ronda, J. I. ; Menéndez, J. M.: Extraction of 3D Structure from Video Sequences. In: García, N. (Hrsg.) ; Martínez, J. M. (Hrsg.) ; Salgado, L. (Hrsg.): *VLBV* Bd. 2849. Springer, 2003, S. 314–322

[85] Jebara, T. ; Azarbayejani, A. ; Pentland, A. : 3D Structure from 2D Motion. In: *IEEE Signal Processing Magazine* 16 (1999), Nr. 3, S. 66–84

[86] Jiang, X. ; Bunke, H. : *Dreidimensionales Computersehen*. Springer, Berlin, 1996

[87] Jiang, X. ; Lambers, M. : DIBR-Based 3D Videos Using Non-Video-Rate Range Image Stream. In: *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*. Toronto, Ontario, Kanada, 2006, S. 1873–1876

[88] Jiang, X. ; Lambers, M. : Synthesis of Stereoscopic 3D Videos by Limited Resources of Range Images. In: *Proc. of International Conference on Pattern Recognition (ICPR)*. Hong Kong, 2006, S. 1220–1224

[89] Jin, H. ; Favaro, P. ; Soatto, S. : Real-Time Feature Tracking and Outlier Rejection with Changes in Illumination. In: *ICCV*, 2001, S. 684–689

[90] Jones, G. A.: Constraint, Optimisation and Hierarchy: Reviewing Stereoscopic Correspondence of Complex Features. In: *Computer Vision and Image Understanding* 65 (1997), Nr. 1, S. 57–78

[91] Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems. In: *Transactions of the ASME–Journal of Basic Engineering* 82 (1960), Nr. Series D, S. 35–45

[92] Kang, S. ; Szeliski, R. ; Chai, J. : Handling Oclusions in Dense Multi-view Stereo / Microsoft Research. Redmond, WA, September 2001 (MSR-TR-2001-80). – Technical Report

[93] Kim, D. ; Min, D. ; Sohn, K. : A Stereoscopic Video Generation Method Using Stereoscopic Display Characterization and Motion Analysis. In: *IEEE Transactions on Broadcasting, Accepted for future publication* (2008)

[94] Kim, J.-H. ; Sikora, T. : Robust Anisotropic Disparity Estimation with Perceptual Maximum Variation Modeling. In: *IEEE International Conference on Image Processing (ICIP)*. Atlanta, GA, USA, October 2006

[95] Kim, K. T. ; Siegel, M. ; Son, J. Y.: Synthesis of a High-Resolution 3D Stereoscopic Image Pair from a High-Resolution Monoscopic Image and a Low-Resolution Depth Map. In: *Proc. of SPIE: Stereoscopic Displays and Applications IX*. San José, CA, USA, 1998

[96] Kitchen, L. ; Rosenfeld, A. : Gray-Level Corner Detection. In: *Pattern Recognition Letters* (1982), S. 95–102

[97] Klette, R. ; Kozera, R. ; Schlüns, K. : Shape from Shading and Photometric Stereo Methods. University of Auckland, New Zealand, 1998 (CITR-TR-20). – Forschungsbericht. – 13 S

[98] Klinke, R. ; Silbernagl, S. : *Lehrbuch der Physiologie*. Stuttgart, New York : Thieme Verlag, 1994

[99] Knorr, S. ; Imre, E. ; Alatan, A. A. ; Sikora, T. : A Geometric Segmentation Approach for the 3D Reconstruction of Dynamic Scenes in 2D Video Sequences. In: *EUSIPCO*. Florence, Italy, September 2006

[100] Knorr, S. ; Kunter, M. ; Sikora, T. : Super-Resolution Stereo- and Multi-View Synthesis from Monocular Video Sequences. In: *3-D Digital Imaging and Modeling (3DIM)*. Montréal, Québec, Canada, August 2007

[101] Knorr, S. ; Sikora, T. : An Image-based Rendering (IBR) Approach for Realistic Stereo View Synthesis of TV Broadcast Based on Structure From Motion. In: *IEEE International Conference on Image Processing (ICIP)*. San Antonio, Texas, USA, September 2007

[102] Knorr, S. ; Smolic, A. ; Sikora, T. : From 2D- to Stereo- to Multi-view Video. In: *3DTV-Conference (3DTV-CON)*. Kos Island, Greece, May 2007

[103] Konrad, J. ; Halle, M. : 3-D Displays and Signal Processing. In: *Signal Processing Magazine, IEEE* 24 (Nov. 2007), Nr. 6, S. 97–111

[104] Laurentini, A. : The Visual Hull Concept for Silhouette-Based Image Understanding. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16 (1994), February, Nr. 2, S. 150–162

[105] Lee, S. ; Feng, D. ; Gooch, B. : Perception-Based Construction of 3D Models from Line Drawings. In: *Symposium on Interactive 3D Graphics and Games*, 2008

[106] Levoy, M. ; Hanrahan, P. : Light Field Rendering. In: *SIGGRAPH: Proc. of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA : ACM, 1996, S. 31–42

[107] Lhuillier, M. ; Quan, L. : Robust Dense Matching Using Local and Global Geometric Constraints. In: *Proc. of International Conference on Pattern Recognition (ICPR), Barcelona, Spain* Bd. 1, 2000. – Piero Zamperoni Best Student Paper Award, S. 968–972

[108] Lipton, L. : *Foundations of the Stereoscopic Cinema – A Study in Depth*. Simon & Schuster, 1982

[109] Lipton, L. : *The CrystalEyes Handbook*. San Rafael, CA, USA : StereoGraphics Corp., 1991

[110] Liu, Y. ; Chen, G. ; Max, N. ; Hoffsetz, C. ; McGuinness, P. : Visual Hull Rendering with Multi-view Stereo Refinement. In: *Proc. of International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)* Bd. 12. Plzen - Bory, Czech Republic, February 2004

[111] Liu, Y.-F. : *A Unified Approach to Image Focus and Defocus Analysis*, State University of New York at Stony Brook, Diss., 1998

[112] Longuet-Higgins, H. : A Computer Algorithm for Reconstructing a Scene from Two Projections. In: *Nature* 293 (1981), September, S. 133–135

[113] Loop, C. T. ; Zhang, Z. : Computing Rectifying Homographies for Stereo Vision. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* Bd. 1. Fort Collins, Colorado, USA, June 1999, S. 1125–1131

[114] Lourakis, M. ; Argyros, A. ; Orphanoudakis, S. : Independent 3D Motion Detection Using Residual Parallax Normal Flow Fields. In: *Proc. of International Conference on Computer Vision (ICCV)*. Bombay, India, 1998

[115] Lourakis, M. ; Argyros, A. : The Design and Implementation of a Generic Sparse Bundle Adjustment Software Package Based on the Levenberg-Marquardt Algorithm / Institute of Computer Science - FORTH. Heraklion, Crete, Greece, Aug. 2004 (340). – Forschungsbericht. Erhältlich unter http://www.ics.forth.gr/ lourakis/sba+

[116] Lucas, B. D. ; Kanade, T. : An Iterative Image Registration Technique with an Application to Stereo Vision. In: *Image Understanding Workshop*, 1981, S. 121–139

[117] Luong, Q.-T. ; Faugeras, O. D.: The Fundamental Matrix – Theory, Algorithms, and Stability Analysis. In: *International Journal of Computer Vision* 17 (1996), Nr. 1, S. 43–75

[118] Luong, Q.-T. ; Faugeras, O. D.: Self-Calibration of a Moving Camera from Point Correspondences and Fundamental Matrices. In: *International Journal on Computer Vision* 22 (1997), Nr. 3, S. 261–289

[119] Ma, Y. ; Soatto, S. ; Košecká, J. ; Sastry, S. : *An Invitation to 3D Vision*. Springer, 2004

[120] Mancini, A. : *Disparity Estimation and Intermediate View Reconstruction for Novel Applications in Stereoscopic Video*, McGill University, Masterarbeit, February 1998

[121] Mark, W. R.: *Post-Rendering 3D Image Warping: Visibility, Reconstruction, and Performance for Depth-Image Warping*, University of North Carolina at Chapel Hill, Diss., 1998

[122] Matsumoto, Y. ; Terasaki, H. ; Sugimoto, K. ; Arakawa, T. : Conversion System of Monocular Image Sequence to Stereo Using Motion Parallax. In: *Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems III*. San Jose, CA, USA, 1996

[123] McMillan, L. : *An Image-Based Approach to Three-Dimensional Computer Graphics*, University of North Carolina at Chapel Hill, Diss., April 1997

[124] Medioni, G. ; Yasumuto, Y. : Corner Ddetection and Curve Representation Using Cubic B-Spline. In: *IEEE International Conference on Robotics and Automation*, 1990, S. 764–769

[125] Meesters, L. M. J. ; Ijsselstein, W. A. ; Seuntiens, P. J. H.: A Survey of Perceptual Evaluations and Requirements of Three-Dimensional TV. In: *IEEE Transactions on Circuits and Systems for Video Technology* 14 (2004), March, S. 381–391

[126] Mendonca, P. R. S. ; Cipolla, R. : A Simple Technique for Self-Calibration. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999

[127] Moré, J. J.: The Levenberg-Marquardt Algorithm: Implementation and Theory. In: Watson, G. A. (Hrsg.): *Numerical Analysis*. Springer Verlag, 1977 (Lecture Notes in Mathematics), S. 105–116

[128] Moravec, H. P.: Visual Mapping by a Robot Rover. In: *International Joint Conference on Artificial Intelligence*, 1979, S. 598–600

[129] Moustakas, K. ; Tzovaras, D. ; Strintzis, M. : Stereoscopic Video Generation Based on Efficient Layered Structure and Motion Estimation from a Monoscopic Image Sequence. In: *Circuits and Systems for Video Technology, IEEE Transactions on* 15 (Aug. 2005), Nr. 8, S. 1065–1073

[130] Müller, P. ; Zeng, G. ; Wonka, P. ; Gool, L. J. V.: Image-based Procedural Modeling of Facades. In: *ACM Transactions on Graphics* 26 (2007), Nr. 3, S. 85

[131] Mulligan, J. ; Isler, V. ; Daniilidis, K. : Trinocular Stereo: A Real-Time Algorithm and its Evaluation. In: *International Journal of Computer Vision* 47 (2002), April, Nr. 1-3, S. 51–61

[132] Myatt, D. ; Torr, P. ; Nasuto, S. ; Bishop, J. ; Craddock, R. : Napsac: High Noise, High Dimensional Robust Estimation – It's in the Bag, 2002, S. 458–467

[133] Oram, D. : Rectification for Any Epipolar Geometry. In: *British Machine Vision Conference (BMVC)*, 2001

[134] Papadimitriou, D. ; Dennis, T. : Epipolar Line Estimation and Rectification for Stereo Image Pairs. In: *IEEE Transactions on Image Processing* 5 (1996), Nr. 4, S. 672–676

[135] Parida, L. ; Geiger, D. ; Hummel, R. : Junctions: Detection, Classification, and Reconstruction. In: *IEEE Transactions on Patern Analysis and Machine Intelligence* 20 (1998), Nr. 7, S. 687–698

[136] Park, S. C. ; Park, M. K. ; Kang, M. G.: Super-Resolution Image Reconstruction: A Technical Overview. In: *IEEE Signal Processing Magazine* 20 (2003), May, Nr. 3, S. 21–36

[137] Parodi, P. ; Piccioli, G. : 3D Shape Reconstruction by Using Vanishing Points. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18 (February 1996), Nr. 2, S. 211–217

[138] Pastoor, S. : 3D-Television: A Survey of Recent Research Results on Subjective Requirements. In: *Signal Processing: Image Communication* 4 (1991), November, Nr. 1, S. 21–32

[139] Pereira, F. ; Ebrahimi, T. : *The MPEG-4 Book*. Prentice Hall, 2002

[140] Pollefeys, M. . *Visual 3D Modeling from Images*. Tutorial notes, tutorial organized in conjunction with 3DPVT 2004, Thessaloniki, Greece. September 2004

[141] Pollefeys, M. ; Gool, L. ; Vergauwen, M. ; Cornelis, K. ; Verbiest, F. ; Tops, J. : Video-to-3D. In: *Photogrammetric Computer Vision (PCV), International Archive of Photogrammetry and Remote Sensing*, 2002, S. 252–258

[142] Pollefeys, M. ; Gool, L. V. ; Vergauwen, M. ; Verbiest, F. ; Cornelis, K. ; Tops, J. ; Koch, R. : Visual Modeling with a Hand-held Camera. In: *International Journal of Computer Vision* 59 (2004), Nr. 3, S. 207–232

[143] Pollefeys, M. ; Koch, R. ; Gool, L. V.: A Simple and Efficient Rectification Method for General Motion. In: *Proc. of International Conference on Computer Vision (ICCV)*. Corfu (Greece), 1999, S. 496–501

[144] Pollefeys, M. ; Koch, R. ; Vergauwen, M. ; Gool, L. V.: Automated Reconstruction of 3D Scenes from Sequences of Images. In: *Isprs Journal Of Photogrammetry And Remote Sensing* 4 (2000), Nr. 55, S. 251–267

[145] Pollefeys, M. ; Gool, L. V.: A Stratified Approach to Metric Self-Calibration. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington, DC, USA, 1997, S. 407

[146] Prados, E. ; Camilli, F. ; Faugeras, O. : A Unifying and Rigorous Shape From Shading Method Adapted to Realistic Data and Applications. In: *Journal of Mathematical Imaging and Vision* 25 (2006), Nr. 3, S. 307–328

[147] Pritchett, P. ; Zisserman, A. : Wide Baseline Stereo Matching. In: *ICCV*, 1998, S. 754–760

[148] Rehg, J. ; Witkin, A. : Visual Tracking with Deformation Models. In: *IEEE International Conference on Robotics and Automation*, 1991, S. 844–850

[149] Rohr, K. : Recognizing Corners by Fitting Parametric Models. In: *International Journal of Computer Vision* 9 (1992), Nr. 3, S. 213–230

[150] Rojo, J. ; Rojas, R. ; Gunnarsson, K. ; Simon, M. ; Wiesel, F. ; Ruff, F. ; Wolter, L. ; Zilly, F. ; Santraè, N. ; Ganjineh, T. ; Sarkohi, A. ; Ulbrich, F. ; Latotzky, D. ; Jankovic, B. ; Hohl, G. ; Wisspeintner, T. ; May, S. ; Pervölz, K. ; Nowak, W. ; Maurelli, F. ; Dröschel, D. : Spirit of Berlin: An Autonomous Car for the DARPA Urban Challenge – Hardware and Software Architecture. 2007. – Forschungsbericht. – 25 S

[151] Ross, J. : Stereopsis by Binocular Delay. In: *Nature* 2 (1974), Nr. 248, S. 363–364

[152] Rotem, E. ; Wolowelsky, K. ; Pelz, D. : Automatic Video to Stereoscopic Video Conversion. In: *Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems XII*. San Jose, CA, USA, March 2005

[153] Rousseeuw, P. ; Leroy, A. : *Robust Regression and Outlier Detection*. Wiley & Sons, New York, 1987

[154] Roy, S. ; Meunier, J. ; Cox, I. : Cylindrical Rectification to Minimize Epipolar Distortion. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997, S. 393–399

[155] Saxena, A. ; Schulte, J. ; Ng, A. Y.: Depth Estimation using Monocular and Stereo Cues. In: *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*. Hyderabad, India, January 2007

[156] Scharstein, D. ; Szeliski, R. : A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. 2001 (Microsoft Research Technical Report MSR-TR-2001-81). – Forschungsbericht. – 35 S

[157] Schindler, K. ; Suter, D. : Two-view Multibody Structure-and-Motion with Outliers through Model Selection. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (June 2006), Nr. 6, S. 983–995

[158] Schmid, C. ; Mohr, R. ; Bauckhage, C. : Evaluation of Interest Point Detectors. In: *International Journal of Computer Vision* 37 (2000), Nr. 2, S. 151–172

[159] Schreer, . ; Kauff, P. : An Immersive 3D Videoconferencing System Based on a Shared Virtual Table Environment. In: *Proc. of International Conference on Media Futures*, 2001

[160] Schreer, O. : *Stereoanalyse und Bildsynthese*. Springer-Verlag, Berlin, Heidelberg, New York, 2005

[161] Schreer, O. ; Kauf, P. ; Sikora, T. : *3D Videocommunication - Algorithms, concepts and real-time systems in human centered communication*. John Wiley & Sons, Ltd, 2005

[162] Schwarz, H. R. ; Köckler, N. : *Numerische Mathematik*. 6. B. G. Teubner Verlag, 2006

[163] Sequeira, V. ; Ng, K. ; E. Wolfart, J. G. M. G. ; Hogg, D. : Automated Reconstruction of 3D Models from Real Environments. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 54 (1999), February, Nr. 1, S. 1–22

[164] Shade, J. ; Gortler, S. ; wei He, L. ; Szeliski, R. : Layered Depth Images. In: *SIGGRAPH '98: Proc. of 25th Annual Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA : ACM, 1998, S. 231–242

[165] Shi, J. ; Tomasi, C. : Good Features to Track. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994, S. 593–600

[166] Silbernagl, S. ; Despopoulos, A. : *Taschenatlas der Physiologie*. Stuttgart, New York : Thieme Verlag, 1983

[167] Smith, S. M. ; Brady, J. M.: SUSAN - A New Approach to Low Level Image Processing. In: *International Journal of Computer Vision* 23 (1997), Nr. 1, S. 45–78

[168] Stoykova, E. ; Alatan, A. A. ; Benzie, P. ; Grammalidis, N. ; Malassiotis, S. ; Ostermann, J. ; Piekh, S. ; Sainov, V. ; Theobalt, C. ; Thevar, T. ; Zabulis, X. : 3-D Time-Varying Scene Capture Technologies: A Survey. In: *IEEE Transactions on Circuits and Systems for Video Technology* 17 (Nov. 2007), Nr. 11, S. 1568–1586

[169] Strecha, C. ; Van Gool, L. : PDE-based Multi-view Depth Estimation. In: *Proc. of International Symposium on 3D Data Processing, Visualization and Transmission* (2002), S. 416–425

[170] Sturm, P. ; Cheng, Z. L. ; Chen, P. C. Y. ; Poo, A. N.: Focal Length Calibration from Two Views: Method and Analysis of Singular Cases. In: *Computer Vision and Image Understanding* 99 (2005), Nr. 1, S. 58–95

[171] Sturm, P. ; Triggs, B. : A Factorization Based Algorithm for Multi-Image Projective Structure and Motion. In: *Proc. of European Conference on Computer Vision (ECCV)*. Cambridge, UK, 1996, S. 709–720

[172] Subbarao, M. ; Choi, T. : Accurate Recovery of Three-Dimensional Shape from Image Focus. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17 (1995), Nr. 3, S. 266–274

[173] Suehiro, K. ; Nakamura, H. ; Yamada, K. ; Nakamura, S. ; Sugahara, T. : Integral 3D Imaging System Using Monocular 2D Video and Depth Data. In: *Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems XII* Bd. 5664. San José, CA, USA, 2005, S. 230–240

[174] Szeliski, R. ; Kang, S. B.: Recovering 3D Shape and Motion from Image Streams Using Non-Linear Least-Squares. In: *Journal of Visual Communication and Image Representation* 5 (1994), March, Nr. 1, S. 10–28

[175] Tam, W. J. ; Speranza, F. ; Zhang, L. ; Renaud, R. ; Chan, J. ; Vazquez, C. : Depth Image Based Rendering for Multiview Stereoscopic Displays: Role of Information at Object Boundaries. In: *Proc. of SPIE: Three-Dimensional TV, Video, and Display IV* Bd. 6016, 2005, S. 75–85

[176] Tekalp, A. M. ; Kurutepe, E. ; Civanlar, M. R.: 3DTV over IP. In: *IEEE Signal Processing Magazine* 24 (2007), November, Nr. 6, S. 77–87

[177] Thormaehlen, T. ; Broszio, H. ; Weissenfeld, A. : Keyframe Selection for Camera Motion and Structure Estimation from Multiple Views. In: *Proc. of European Converence on Computer Vision (ECCV)*, 2004, S. 523–535

[178] Tola, E. ; Knorr, S. ; Imre, E. ; Alatan, A. A. ; Sikora, T. : Structure from Motion in Dynamic Scenes with Multiple Motions. In: *2nd Workshop On Immersive Communication and Broadcast Systems (ICOB '05)*. Berlin, Germany, October 2005

[179] Tomasi, C. ; Kanade, T. : Shape and Motion from Image Streams: A Factorization Method. In: *Journal of Computer Vision* 9 (1992), Nr. 2, S. 137–154

[180] Tomasi, C. ; Kanade, T. : Detection and Tracking of Point Features / Carnegie Mellon University. 1991 (CMU-CS-91-132). – Forschungsbericht

[181] Torr, P. : An Assessment of Information Criteria for Motion Model Selection. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997, S. 47–53

[182] Torr, P. : Geometric Motion Segmentation and Model Selection. In: *Phil. Trans. Royal Society of London A* 356 (1998), May, Nr. 1740, S. 1321–1340

[183] Torr, P. : Geometric Motion Segmentation and Model Selection. In: *Royal Society of London Philosophical Transactions Series A* 356 (1998), May, S. 1321–1340

[184] Torr, P. : Bayesian Model Estimation and Selection for Epipolar Geometry and Generic Manifold Fitting. In: *International Journal of Computer Vision* 50 (2002), S. 35–61

[185] Torr, P. ; Fitzgibbon, A. ; Zisserman, A. : The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Images. In: *International Journal of Computer Vision* 32 (1999), August, Nr. 1, S. 47–53

[186] Torr, P. ; Zisserman, A. : MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. In: *Computer Vision and Image Understanding* 78 (2000), S. 138–156

[187] Treuilett, S. ; Albouy, B. ; Lucas, Y. : Finding Two Optimal Positions of a Handheld Camera for the Best Object Reconstruction. In: *3DTV-Conference (3DTV-CON)*. Kos Island, Greece, May 2007

[188] Triggs, B. ; McLauchlan, P. ; Hartley, R. ; Fitzgibbon, A. : Bundle Adjustment – A Modern Synthesis. In: Triggs, W. (Hrsg.) ; Zisserman, A. (Hrsg.) ; Szeliski, R. (Hrsg.): *Vision Algorithms: Theory and Practice*. Springer Verlag, 2000 (LNCS), S. 298–375

[189] Trucco, E. ; Plakas, C. ; Brandenburg, N. ; Kauff, P. ; Karl, M. ; Schreer, O. : Real-Time Disparity Analysis for Immersive 3-D Teleconferencing by Hybrid Recursive Matching and Census Transform. In: *Proc. of Workshop on Video Registration, IEEE International Conference on Computer Vision (ICCV)*. Vancouver, Canada, July 2001

[190] Trujillo, M. ; Izquierdo, E. : A Robust Correlation Measure for Correspondence Estimation. In: *2nd International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT)*, 2004

[191] Vergauwen, M. ; Pollefeys, M. ; Gool, L. V.: A Stereo Vision System for Support of Planetary Surface Exploration. In: *Machine Vision and Applications* 14 (2003), Nr. 1, S. 5–14

[192] Vidal, R. ; Ma, Y. ; Soatto, S. ; Sastry, S. : Two-View Multibody Structure from Motion. In: *International Journal of Computer Vision (IJCV), special issue on Dynamic Vision* 68 (2006), June, Nr. 1, S. 7–25

[193] Walter, T. : *Robuste 3D Szenenanalyse in Videosequenzen mit Hilfe von zeitlicher Segmentierung*, Technische Universität Berlin, Diplomarbeit, November 2006

[194] Wei, T.-C. : *Three Dimensional Machine Vision Using Image Defocus*, State University of New York at Stony Brook, Diss., 1994

[195] Wolf, L. ; Shashua, A. : Two-body Segmentation from Two Perspective Views. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Kauai Marriott, Hawaii, USA, December

[196] Woods, A. ; Docherty, T. ; Koch, R. : Image Distortions in Stereoscopic Video Systems. In: *Proc. of SPIE: Stereoscopic Displays and Applications IV*. San Jose, CA, USA, 1993, S. 36–48

[197] Würmlin, S. ; Lamboray, E. ; Waschbüsch, M. ; Kaufmann, P. ; Smolic, A. ; Gross, M. : Image-space Free-viewpoint Video. In: *Proc. of Vision, Modeling, Visualization (VMV)*, 2005, S. 453–460

[198] Zhang, L. ; Tam, J. ; Wang, D. : Stereoscopic Image Generation Based on Depth Images. In: *IEEE International Conference on Image Processing (ICIP)*. Singapore, 2004

[199] Zhang, L. ; Tam, W. : Stereoscopic Image Generation Based on Depth Images for 3D TV. In: *IEEE Transactions on Broadcasting* 51 (2005), Nr. 2, S. 191–199

[200] Zhang, R. ; Tsai, P.-S. ; Cryer, J. E. ; Shah, M. : Shape from Shading: A Survey. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21 (1999), Nr. 8, S. 690–706

[201] Zhang, Z. ; Deriche, R. ; Faugeras, O. D. ; Luong, Q.-T. : A Robust Technique for Matching two Uncalibrated Images through the Recovery of the Unknown Epipolar Geometry. In: *Artificial Intelligence* 78 (1995), Nr. 1–2, S. 87–119

[202] Zhang, Z. : A Flexible New Technique for Camera Calibration / Microsoft Research. Redmond, WA, (MSR-TR-98-71). – Forschungsbericht

[203] Zhang, Z. : A Stereovision System for a Planetary Rover: Calibration, Correlation, Registration, and Fusion. In: *Machine Vision and Applications* 10 (1997), Nr. 1, S. 27–34

[204] Zhao, X. : *Optimierung von 3D Rekonstruktionsergebnissen mit Hilfe globaler Verarbeitungsmethoden*, Technische Universität Berlin, Studienarbeit, June 2007

[205] Zitnick, C. ; Kanade, T. : A Cooperative Algorithm for Stereo Matching and Occlusion Detection. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (2000), Nr. 7, S. 675–684

[206] Zitnick, C. L. ; Kang, S. B. ; Uyttendaele, M. ; Winder, S. ; Szeliski, R. : High-Quality Video View Interpolation Using a Layered Representation. In: *ACM Transactions on Graphics* 23 (2004), Nr. 3, S. 600–608