Automatische Generierung von 3D-Modellen mittels Sequenzen hochauflösender Bildtripel

vorgelegt von

Diplom-Ingenieur

Matthias Heinrichs

aus Berlin

von der Fakultät IV - Elektrotechnik und Informatik

der Technischen Universität Berlin

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften

- Dr.-Ing. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Marc Alexa
Berichter:	Prof. Dr. Uwe Stilla (Technische Universität München)
Berichter:	Prof. Dr. Olaf Hellwich

Tag der wissenschaftlichen Aussprache: 10. Dezember 2010

Berlin 2011

D 83

Für Felix

Zusammenfassung

Die Rekonstruktion einer räumlichen Szene aus Bildaufnahmen kann automatisch erfolgen, wenn Kameraposition und -orientierung sowie korrespondierende Bildpunkte ohne Eingreifen eines Operators berechnet werden können. Da die Aufnahme einer Szene in der Regel zügig erledigt ist und daraus eine Fülle von Daten entsteht, ist es besonders wichtig, dass diese Daten vollautomatisch ausgewertet werden können, um eine hochgenaue 3D-Rekonstruktion für die weitere Verarbeitung ohne viel Aufwand zu ermöglichen. Durch die Vielzahl von möglichen Schwierigkeiten bei der Aufnahme muss ein solches System ferner besonders robust und fehlertolerant sein, um möglichst vielseitig eingesetzt werden zu können.

Da bei einem Bildpaar prinzipiell nur die in beiden Bildern gleichzeitig sichtbaren Objekte rekonstruiert werden können, müssen für eine vollständige Rekonstruktion des Objektes mehrere Rekonstruktionen aus verschiedenen Ansichten zusammengefügt werden. Dies ist nur möglich, wenn die Kamerapositionen der verschiedenen Ansichten bekannt sind. Für eine hochgenaue Rekonstruktion müssen daher die Aufnahmepositionen für jedes Bild bekannt sein und sehr viele Bilder verarbeitet werden können, um auch geometrisch anspruchsvolle Objekte mit vielen Verdeckungen und großen Dimensionen rekonstruieren zu können.

Diese Arbeit erweitert, verfeinert und passt aktuelle Forschungsergebnisse der Themenbereiche Merkmalsextraktion, relative Orientierung, Rektifizierung und Korrespondenzanalyse aneinander an, um ein einheitliches, aufeinander abgestimmtes System zu integrieren. Dabei werden sowohl Videodaten für eine einfache Verfolgung der Szeneninhalte als auch Stereobildaufnahmen von verschiedenen Videokameras mit einem geeigneten Kamerabstand für eine genaue Triangulation verwendet, wobei sich die jeweiligen Vorteile der Rekonstruktionstechniken ergänzen und zu einem funktionierenden Gesamtsystem führen. Das vorgestellte System verwendet dabei drei synchronisierte Videoströme von drei Kameras, die auf einem Rahmen zueinander fest montiert sind. Als Voraussetzung müssen lediglich die intrinsische Kalibrierung und radiale Verzeichnung der Kameralinsen bekannt sowie die Synchronität der drei Videoströme gewährleistet sein.

Die Grundvoraussetzung für eine genaue Rekonstruktion ist eine stabile und akkurate Zuordnung von Punktmerkmalen. Dazu werden vorhandene Interestoperatoren und Merkmalsdeskriptoren so kombiniert und die Subpixelbestimmung so verfeinert, dass die daraus resultierenden Merkmale besonders gut für die Auswertung von Videos geeignet sind.

Für die robuste Zuordnung dieser Merkmale wird eine neue Technik vorgestellt, anhand derer aus den Beziehungen der drei Kameras zueinander und drei aufeinander folgenden Bildern jeder dieser Kameras Bedingungen abgeleitet werden können, um nahezu sämtliche Fehlzuordnungen herauszufiltern. Ein weiterer Vorteil ist die gleichzeitige Zuordnung der Merkmale sowohl im zeitlichen Verlauf eines Videostroms als auch zwischen den Bildern der drei Kameras. Dadurch entstehen räumliche/zeitliche Zuordnungen und es können hilfreiche Bedingungen eingeführt werden, die durch die feste Anordnung der Kameras auf dem Rahmen entstehen. Ferner ist es möglich, in gewissen Grenzen die Position von Punktmerkmalen vorherzusagen und verlorene oder verdeckte Korrespondenzen zu reparieren oder wiederzufinden.

Die Rekonstruktion des Kamerapfades mittels Bildkorrespondenzen und relativer Orientierung wurde auf das vorhandene Dreikamerasystem erweitert. Wegen der räumlich/zeitlichen Korrespondenzen können die Mehrdeutigkeiten in der Pfadbestimmung auf ein Minimum reduziert werden. Gleichzeitig wird die Bestimmung eines global gültigen Maßstabs durch den Rahmen vereinfacht und es kann gezeigt werden, dass der Maßstabsfehler auch nach Tausenden von Bildern nicht zunimmt.

Für eine dichte Zuordnung der Bildpunkte wird vor der eigendlichen Korrespondenzsuche häufig eine Bildrektifizierung eingefügt, die im Stereobildfall die Bildzeilen so anordnet, dass sie mit den Epipolarlinien korrespondieren. In dieser Arbeit wird ein lineares Verfahren vorgestellt, anhand dessen drei Bilder rektifiziert werden können, so dass die Bildzeilen und -spalten eines Referenzbildes jeweils mit den Epipolargraden von einem der zwei anderen Bilder korrespondieren. Diese Vorverarbeitung ermöglicht es, die drei Bilder bei der Korrespondenzsuche so zu untersuchen, dass im Vergleich zur Zweibildanalyse kaum Mehraufwand entsteht.

Diese Rektifizierung bildet die Basis für eine automatische und dichte Zuordnung von Bildpunkten hochauflösender Bilder. Es werden in dieser Arbeit Erweiterungen für ein robustes Zuordnungsverfahren vorgestellt, um in drei Bildern gleichzeitig Korrespondenzen zu finden und über den Aufbau von Bildpyramiden die Zuordnungsqualität zu verbessern. Des weiteren wird ein Verfahren zur Subpixelbestimmung beschrieben, um Informationen, die aus den Aliaseffekten des Bildrasters entstehen, so genau wie möglich zu integrieren.

Die Funktion des Systems wird an realen Daten demonstriert und die Genauigkeit der Ergebnisse mit etablierten Messmethoden bestätigt. Es kann gezeigt werden, dass auch große Objekte mit sehr hoher Genauigkeit rekonstruiert werden können und dabei auf externe Sensorik wie Beschleunigungsmesser, Gyroskope und GPS-Empfänger verzichtet werden kann.

Abstract

A three-dimensional (3D) scene can be automatically reconstructed from images if camera position and orientation as well as corresponding image points can be determined without the aid of an operator. Recording a scene usually takes only little time while generating plenty of data, it is therefore important that the data be processed automatically and without too much effort in order to enable highly precise 3D reconstruction for further use. In addition, due to the multitude of difficulties that could arise during recording, the system must be exceptionally robust and error-tolerant to guarantee maximum applicability.

In an image pair, only those objects visible in both images at the same time can be reconstructed. For complete reconstruction of an object, therefore, several reconstructions from different viewpoints must be assembled, which is only possible if all camera positions are known. Thus, locations and orientations for each image viewpoint must be determined and a great number of images must be processed for highly precise reconstructions even of geometrically sophisticated objects containing many occlusions and covering large dimensions.

In this paper, current research in the fields of feature point extraction, relative orientation, rectification and image registration is expanded, refined and adjusted to create one coherent system. Video data for scene tracking as well as stereo images taken by different video cameras with an appropriate distance between them are used for exact triangulation, the respective advantages of the different reconstruction techniques thus complementing each other and producing an efficient system. For the system presented in this paper, three synchronized video streams taken by three cameras fixed on a frame are used, the only preconditions being that the intrinsic calibration and radial distortion of the camera lenses must be known and the synchrony of the three video streams must be guaranteed.

Precise reconstruction is only possible if feature points are matched reliably and accurately. In order to achieve this, exsiting interest operators and feature descriptors are combined and sub-pixel determination is refined so that the resulting feature points are perfectly fitted for video evaluation.

For the reliable matching of those features, a new technique is presented, by means of which conditions can be derived from the relation between the three cameras and from three consecutive images taken by each of these cameras to filter almost all mismatches. Furthermore, the system allows synchronous feature matching, taking place chronologically along the video stream as well as between the images taken by the three cameras. This generates spatial and temporal matches and useful conditions deriving from the fixed arrangement of the cameras on the frame can be introduced into the system. Thus, within certain limits, the position of feature points becomes predictable, lost correspondences can be found and errors resulting from occlusions can be fixed.

Camera path reconstruction through image correspondences and relative orientation was expanded to fit the existing trifocal system. Ambiguities in path reconstruction can be minimized due to spatial/temporal correspondences. At the same time, the frame facilitates the determination of a global scale, and evidence shows that scale errors do not increase even after thousands of images.

In order to achieve dense stereo view correspondences, an image rectification is often conducted before the actual correspondence search. Thus, in the stereo normal case, image rows are arranged in such a way that they correspond with the epipolar lines. In this paper, a linear method is presented to rectify three images so that the image rows and columns of a reference image correspond with the respective epipolar lines of the two other images. When this method is applied, a correspondence search from three images requires almost no additional effort compared to a two-image analysis.

This rectification forms the basis for the automatic generation of dense stereo correspondences in high-resolution images. Expansions for a robust matching method are presented in this paper to simultaneously find correspondences in three images and improve the matching quality by building image pyramids. Furthermore, a method for sub-pixel determination is described so that information deriving from the picture raster's alias effects can be integrated as precisely as possible.

The system's functionality is demonstrated using real data, and the exactitude of the results is confirmed using well-established measuring methods. It is shown that even large objects can be precisely reconstructed without the aid of external sensor technology such as accelerometers, gyroscopes or GPS receivers.

Inhaltsverzeichnis

Teil I
Einleitung und Motivation

1 Thema und Problemstellung

Die Bestimmung dreidimensionaler (3D) Raumpositionen mittels stereoskopischer Bildauswertung ist der Kernbereich der klassischen Photogrammetrie und ein Hauptthema der modernen Computer Vision. Die aus den entsprechenden Forschungen resultierenden Rekonstruktionssysteme werden in zwei Kategorien unterteilt: aktive und passive Systeme. Bei aktiven Systemen wird ein definierter Licht-, Radar- oder Laserstrahl von einem Sender, der Teil des Systems ist, ausgesendet und von einem Sensorelement registriert. Mittels geeigneter Methodik wird somit aus Position, Phasenversatz oder Intensität ein Tiefenwert bestimmt. Bekannte aktive Systeme sind Laserscanner, Rekonstuktion durch strukturierte Beleuchtung, Ultraschall oder Radar.

Bei passiven Systemen hingegen wird das Objekt ausschließlich mit Hilfe der Umgebungsbeleuchtung und eventuell der Eigenstrahlung auf das Sensorelement abgebildet, wodurch der Vorteil entsteht, dass keine zusätzliche Licht- oder elektromagnetische Strahlung verwendet wird, die aus diversen Gründen nicht erwünscht oder möglich sein kann. Da nur ein Sensor benötigt wird, können Systeme dieser Klasse sehr einfach aufgebaut sein. Im Minimalfall genügt eine Kamera, die das Objekt fotografiert. Des weiteren kann durch den Verzicht auf zusätzliche Strahlung die natürliche Farbe des Objektes rekonstruiert werden. Allerdings kann auf diese Weise mit nur einer Messposition nicht auf die Objektstruktur geschlossen werden, da die Beleuchtungseigenschaften unbekannt sind. Daher benötigt man für eine Messung immer zwei Abbildungen, die möglichst unter denselben Beleuchtungsverhältnissen aufgenommen wurden.

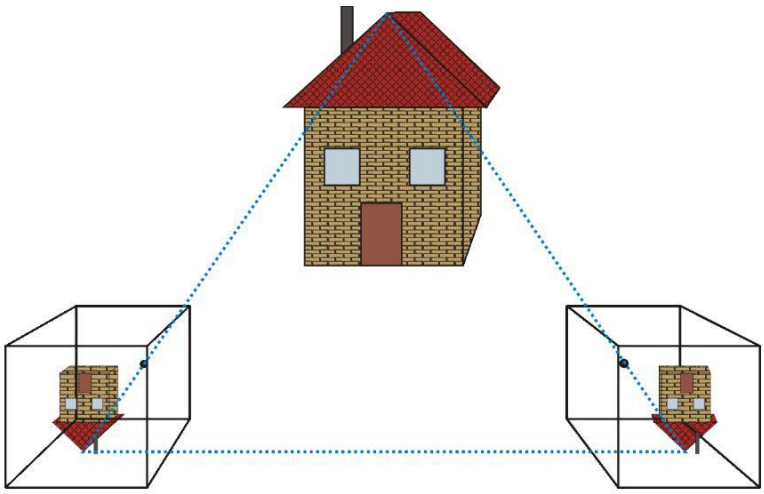


Abbildung 1: Triangulation mittels Lochkamera

Die Rekonstruktion der Oberflächen komplexer Objekte aus Bildern mittels passiver Systeme ist möglich, wenn die Abbildungsgesetze bekannt sind, das Objekt aus zwei verschiedenen Positionen aufgenommen wurde und die korrespondierenden Bildpunkte desselben Objektes zugeordnet wurden. Beim gängigen Lochkameramodell, das bereits von Aristoteles erkannt und im Werk *Problemata physica* [1] erstmalig erwähnt wurde, bilden die Lichtstahlen des Objektes, die durch die Kameraöffnung auf die Bildebene fallen, zusammen mit der Verbindungslinie der Projektionsstellen ein Dreieck (vgl. Abbildung 1). Sind Kameraorientierung und Verbindungsline bekannt, kann das Dreieck und damit der Raumpunkt rekonstruiert werden. Man spricht in diesem Zusammenhang von der Triangulation des Raumpunktes. Verdeckte Teile eines Objektes können auf diesem Wege allerdings nicht

rekonstruiert werden. Für sie müssen zwei weitere Kameraposition gefunden werden, die den verdeckten Bereich abbilden, wobei natürlich eine der bisherigen Kamerapositionen wiederverwendet werden kann.

Die beiden Aufnahmepositionen können auf unterschiedliche Arten entstehen. Zum einen können zwei Kameras die Szene zum selben Zeitpunkt aus unterschiedlichen Positionen aufnehmen. Da die Aufnahmen nur räumlich getrennt sind, wird ein solches Verfahren in dieser Arbeit *räumliches Stereo* genannt. Eine andere Möglichkeit besteht darin, das Objekt mit derselben Kamera nacheinander von unterschiedlichen Positionen aufzunehmen, z.B. durch eine Videosequenz. Dies kann natürlich nur bei starren Objekten durchgeführt werden, da Bewegungen des Objektes nicht immer von der Bewegung der Kamera unterschieden werden können. Da sich die Aufnahmen hierbei nicht nur räumlich, sondern auch zeitlich unterscheiden, wird dieses Verfahren hier *zeitliches Stereo* genannt.

Die Bestimmung einer Raumposition aus Bilddaten gliedert sich thematisch in zwei Teile: Bestimmung der Kamerageometrien und damit der Abbildungsgesetze und Zuordnung der korrespondierenden Objektabbildung. Ob der untersuchte Objektabschnitt überhaupt in beiden Abbildungen sichtbar ist, wird hierbei als Teilproblem der Zuordnung angesehen. Werden diese zwei Teile automatisch von einem Computer bearbeitet, kann das Objekt automatisch rekonstruiert werden, was letztendlich durch die nun mögliche Triangulation zu einem 3D-Modell der Szene führt. Das Objekt wird in diesem Modell durch eine Wolke von Raumkoordinaten beschrieben. Wird es nur von zwei Standpunkten aus aufgenommen, kann lediglich der sichtbare Teil rekonstruiert werden und der den Kameras abgewandte Teil bleibt unbekannt. Man spricht in diesem Fall von 2.5D, da für eine "echte 3D-Rekonstruktion" die Informationen über die verdeckten Teile fehlen. Für eine vollständige Rekonstruktion muss das Objekt aus sehr vielen Richtungen aufgenommen werden. Dafür eignen sich Videodaten, da das Objekt bei einer Videoaufzeichnung "im Vorbeigehen" aus zahlreichen Blickwinkeln aufgenommen wird.

Für die Bestimmung der Kamerageometrie gibt es verschiedene Verfahren. Ein klassischer Ansatz besteht darin, die Kamerageometrie über bekannte Raumpunkte, auch Passpunkte genannt, zu bestimmen und daraus Ausrichtung, Position und Abbildungseigenschaften zu berechnen. Die Kameraposition bezieht sich bei dieser Methode auf das Koordinatensystem der Passpunkte. Durch hochgenaue Sensorik ist es auch möglich, die Bewegungen der Kamera direkt zu messen und zu rekonstruieren. Solche Systeme werden beispielsweise in der Luftbilderstellung eingesetzt. Hierbei beziehen sich alle Kamerapositionen auf die initiale Position und es werden nur die Unterschiede registriert. Eine globale Referenzierung erfolgt hierbei entweder über Registrierung der Bilder mit bekannten Position von Passpunkten oder über Verknüpfung mit GPS- bzw. differentialGPS-Informationen. Ein dritter Ansatz bestimmt die Kameraposition ausschließlich aus der Bildbewegung. Diese relative Orientierung kann ohne aufwändige Sensorik oder Passpunkte die Ausrichtung und Bewegung der Kamera bestimmen, nicht jedoch den Maßstab, da es aus der Bildinformation alleine nicht ersichtlich ist, ob das Objekt 10mm oder 10km entfernt ist.

Die Korrespondenzsuche zwischen zwei Stereobildern ist nur mit aufwändigen Algorithmen zufriedenstellend lösbar. Das Grundproblem ist hierbei, einem Rechner den erforderlichen Kontext zur Verfügung zu stellen, um Mehrdeutigkeiten aufzulösen. Dieser Kontext ist allerdings nur schwer zu definieren und es muss hierbei die Balance zwischen Universalität mit einer höheren Fehlerrate oder Spezialisierung mit einer geringeren Fehlerrate gehalten werden. Bei realen Szenen kommt es fast unvermeidbar zu Fehlzuordnungen, grade bei wiederkehrenden Mustern oder homogenen Flächen. Auch der Mensch mit seinen hoch entwickelten Sehfähigkeiten ist nicht vor Fehlzuordnungen gefeit, was bei optischen Illusionen wie Single Image Random Dot Stereograms (SIRDS) [30] und den darauf basierenden, in den 90er Jahren modernen "Magic Eye"-Büchern bewusst ausgenutzt wird.

Des weiteren gibt es bei der Zuordnung ein prinzipielles Problem, welches in Abbildung 2 illustriert ist: Links liegen zwei Aufnahmepositionen dicht beieinander und die Abbildungen der Kameras sind aufgrund der geringen perspektivischen Verzerrung sehr ähnlich. Eine Zuordnung der Bildinhalte ist hier relativ leicht. Allerdings entsteht durch den geringen Abstand ein sehr schleifender Schnitt im Raum, der nicht genau bestimmbar ist. Vergrößert man den Abstand der Aufnahmepositionen (rechts), ist der Schnitt im Raum besser bestimmbar.

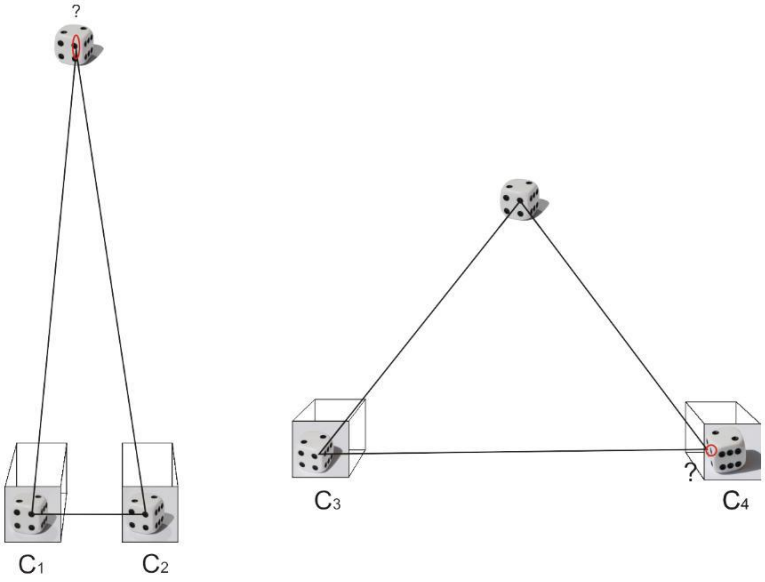


Abbildung 2: Zusammenhang Tiefengenauigkeit und Zuordnungsgenauigkeit

Allerdings nimmt die perspektivische Verzerrung zu und die Zuordnung der Bildinhalte wird schwieriger oder ist gar nicht mehr möglich. Meist wird bei räumlichen Stereoverfahren mit größeren Kameraabständen als beim zeitlichen Stereo mittels Videosequenzen gearbeitet, daher ist die Triangulation beim räumlichen Stereo mathematisch genauer, während Stereoinformationen aus Videodaten wegen der kürzeren Abstände eine geringe Tiefenauflösung haben.

Für eine hochgenaue Rekonstruktion ist es wegen der Komplexität der Materie unabdingbar, die potenziellen Fehlerquellen in jedem Schitt möglichst genau zu modellieren und möglichst gering zu halten. Die Genauigkeit der Rekonstruktion hängt von vielen Faktoren ab. Zunächst müssen die realen Abbildungseigenschaften der Kamera durch die Abbildungsgesetze korrekt beschrieben sein. Dazu gehören der physikalische Strahlengang des projizierten Bildes durch den Linsenapparat sowie die räumliche Position und die Ausrichtung der Kameras. Hierbei können durch Abbildungsunschärfe und Positionierungsungenauigkeiten Fehler entstehen. Die Positionierungsgenauigkeit des Kamerastandortes bei der Aufnahme gehört zu den wichtigsten Faktoren für die Rekonstruktionsqualität. Die klassische Photogrammerie liefert hier mit dem Ansatz der Ausgleichsrechnung ein probates Mittel, um diesen Fehler so klein wie möglich zu halten und gleichmäßig zu verteilen. Ein weiterer Faktor ist die Genauigkeit der Zuordnung des Objektes, die maßgeblich von der Auflösung des projizierten Bildes abhängt. Mit der Auflösung hängen ferner die Objektgröße und der Abstand der Aufnahmepositionen zusammen. Zu guter Letzt muss die Aufnahme sorgfältig geplant werden: Kameraeinstellung, -abstand, Objektgröße, Lichtverhältnisse und Anspruch an die Messergebnisse bedingen bestimmte Parameter wie Belichtungszeit, Basislänge und Fokussierung, die wiederum die Qualität der Rekonstruktion maßgeblich beeinflussen.

2 Zielsetzung und Methodik

Ziel dieser Arbeit ist ein System, das die 3D-Szene vollautomatisch aus Bilddaten rekonstruiert. Dazu sollten die Vorzüge von räumlichen und zeitlichen Stereotechniken kombiniert werden, um die Nachteile der einzelnen Techniken wieder auszugleichen (vgl. Tabelle 1). Es wird gezeigt, dass die räumliche Stereorekonstruktion aus drei Bildern nur unwesentlich aufwändiger ist als die Rekonstruktion aus zwei Bildern, wenn eine entsprechende geometrische Anpassung der Bilder erfolgt. Dabei haben drei Bilder den Vorteil, dass redundante Informationen benutzt werden können, um Fehlzuordnungen zu minimieren. Es wird ein universelles Korrespondenzsuchverfahren vorgestellt, das mit vertretbarem Rechenaufwand automatisch

Kriterium	RS	ZS
Bewegliche Objekte erlaubt	Ja	Nein
Perspektivische Verzerrung	Mittel-Hoch	Niedrig
Gute Triangulation	Ja	Nein
Verdeckung	Hoch	Niedrig
Kameraabstand	Fix	Variabel
Objektgröße beschränkt durch	Bildgröße	Pfadlänge
Bewegungsunschärfe	Kaum	Evtl. stark
Flexibilität	Mittel	Hoch

Tabelle 1: Vergleich von räumlichem Stereo (RS) und zeitlichem Stereo (ZS)

auf Bildtripeln arbeitet und universell einsetzbar ist.

Die Auflösung von Videokameras hat in den letzten Jahren rapide zugenommen, wodurch diese Systeme mittlerweile für hochgenaue 3D-Rekonstruktion geeignet sind. Es wird gezeigt, dass die Bewegung der Kameras allein aus den Bildveränderungen rekonstruiert werden kann. Hierbei ist es besonders wichtig, dass markante Bildpunkte in Videosequenzen präzise lokalisiert und robust verfolgt werden können, um die Kamerabewegung genau zu berechnen.

Die räumlichen und zeitlichen Stereoverfahren werden so integriert, dass das 3D-Modell eines Objektes automatisch nahezu vollständig rekonstruiert werden kann. Dabei wird bewusst auf weitere Sensorik verzichtet, wodurch das System einfacher in Handhabung und Aufbau ist und auch an Orten verwendet werden kann, an denen bestimmte Informationen wie GPS nicht zur Verfügung stehen.

3 Übersicht und Kapitelstruktur

Diese Arbeit ist in vier Teile untergliedert. Nach dieser Einleitung folgt eine ausführliche Beschreibung der theoretischen Grundlagen, wobei auch der Stand der Forschung erklärt wird und die nötigen Fachbegriffe eingeführt sowie die zu lösenden Probleme definiert werden. Dieser Theorieteil umfasst Grundlagen zur digitalen Bildverarbeitung von Videodaten, geometrischen Transfrmationen im Raum, Bestimmung von Punktmerkmalen, Verfolgung dieser Punktmerkmale, Verfahren zur Korrespondenzsuche und die Kalibrierung und Orientierung der Kameras.

Im darauf folgenden Teil wird gezeigt, wie die vorhandenen Techniken integriert und erweitert wurden, um eine automatische Analyse von Bildsequenzen zu ermöglichen. Dieser Teil beginnt mit der Zuordnung und Verfolgung markanter Punktmerkmale und erweitert die Techniken der relativen Orientierung, um eine hochgenaue Pfadschätzung zu erhalten. Die zweite Hälfte dieses Teils beschäftigt sich mit der geometrischen Anpassung der drei Bilder, der Rektifizierung, die die Vorraussetzung für die anschließende räumliche Korrespondenzsuche darstellt. Im Anschluss daran wird das Verfahren zur 3D-Rekonstruktion aus diesen Korrespondenz- und Kameradaten erläutert. Abgeschlossen wird dieser Teil mit der Integration der vorgestellten Module.

Im letzten Teil werden die Ergebnisse der Arbeit zusammengefasst und diskutiert. Abschließend wird ein Ausblick auf Erweiterungen und weiterführende Forschungsmöglichkeiten gegeben.

Teil II
Theoretische Grundlagen

In diesem Kapitel werden die Grundlagen der Forschungshypothese dargestellt. Es beschreibt alle für das System benötigten Gebiete und definiert die verwendeten Begriffe. Ferner befasst es sich mit den grundlegenden Themen der digitalen Bildverarbeitung, Punktmerkmalen, Videotracking, Korrespondenzsuche in Stereobildern und Kamerakalibrierung. Die Namenskonvention in dieser Arbeit ist in Anhang A erläutert.

4 Digitale Bildverarbeitung

Aktuelle, digitale Kamerasysteme nehmen hochauflösende Farbbilder auf. Jeder Farbpunkt wird durch seine drei Grundfarben rot, grün und blau (RGB) beschrieben. Der Sensor, der als Projektionsfläche in der Kamera dient, besitzt ein regelmäßiges Raster an Sensorelementen. Um Farbbilder zu erhalten, werden vor diesen Sensorelementen unterschiedliche Farbfilter angebracht, die vorwiegend nach dem "Bayer-Tile"-Verfahren angeordnet werden (vgl. Abbildung 3).

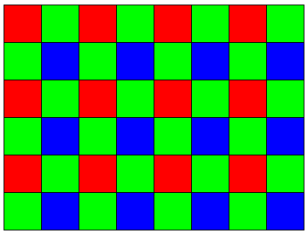


Abbildung 3: Farbfilteranordnung im Bayer-Tile-Schema

Hierbei sind die Farbfilter in einem 2x2-Block im Schema RG-GB oder in einer Verschiebung dieses Schemas über die Sensorfläche verteilt. Der doppelte Grünanteil soll dem menschlichen Sehempfinden Rechnung tragen, das Grüntöne besonders gut differenzieren kann. Allerdings sinkt durch diese Filteranordnung die physikalische Auflösung auf ein Viertel im Rot- und Blau-Kanal und auf die Hälfte im Grünkanal. Eine 4MP-Kamera hat dementsprechend nur jeweils eine Million rote und blaue und zwei Millionen grüne Farbwerte. Das regelmäßige Raster erlaubt jedoch zwischen den Farbtönen zu interpolieren, wodurch die Auflösung mit einem geringen Verlust an Genauigkeit wieder auf die gesamte Rastergröße hochgerechntet wird. Dieses "Demosaicing" versucht, für jeden Farbkanal aus den Intensitätswerten der benachbarten Messungen einen möglichst genauen Intensitätswert zu berechnen oder einen akzeptablen Kompromiss zwischen Qualität und Rechenzeit zu finden. Dabei fließen - je nach Verfahren - Gradientenrichtung, Fenstergröße, Frequenz der Fouriertransformation und Glättungsfunktionen in die Berechnung der Interpolation ein. Gute Übersichten aus gängigen Interpolationsverfahren sind in [41, 17] aufgeführt. Keines der bekannten Verfahren kann jedoch verhindern, dass in bestimmten Fällen an Kanten Farbverfälschungen, insbesondere Moiré-Effekte, entstehen.

4.1 Bildqualität

Die aufgenommenen Farben hängen von Beleuchtung der Szene und Weißabgleich der Kameras ab. Die Grundvoraussetzung beim Vergleich von Bildern ist, dass die Farben nicht stark verfälscht sind. Drei Hauptursachen für Farbverfälschungen sind Rauschen, Beleuchtungsänderungen und fehlerhafte Farbkalibrierung.

Das Verhältnis von Rauschen und Messung wird durch das Signal-Rausch-Verhältnis (SNR) in Dezibel angegeben. Das SNR wird aus dem Verhältnis des mittleren Signalwerts

zum mittleren Rauschwert gebildet, wobei sich die Normierungsfaktoren gegenseitig aufheben. Da die Signalleistung meist sehr viel höher ist als der Rauschwert, wird das Ergebnis logarithmisch aufgetragen:

$$snr = 10ln\left(\frac{\sum S_x}{\sum N_x}\right) \tag{1}$$

Fällt weniger Licht auf den Sensor, nimmt das durchschnittliche Signal ab, während der Rauschanteil häufig konstant bleibt. Dunkle Aufnahmen haben daher i.d.R. ein schlechteres SNR. Eine zu hohe Lichtmenge hingegen führt zu einer Sättigung der Sensorelemente und die Intensität kann nicht korrekt gemessen werden. Daher muss die aufgenommene Lichtmenge, die abhängig von Beleuchtung, Blende und Belichtungszeit ist, möglichst konstant und unterhalb der Sättigungsschwelle bleiben, um die Rekonstruktion nicht durch ein ungleichmäßiges Rauschverhalten zu erschweren.

Die Ausleuchtung eines Objektes ist davon abhängig, in welchem Winkel die Kamera zur Objektoberfläche steht. Das einfallende Licht wird gemäß den Reflexionsgesetzen reflektiert. Dabei gibt es zwei Arten: Totalreflexion und diffuse Reflexion an rauen Oberflächen. Die Totalreflexion reflektiert das einfallende Licht fast vollständig in dem selben Winkel zur Oberflächennormale, in dem es einfällt. Der Betrachter sieht das Spiegelbild auf der Oberfläche. Auf einer diffusen Oberfläche wird die Lichtreflexion durch das Lambertsche Gesetz beschrieben: Die Intensität des reflektierten Lichts nimmt mit dem Cosinus des Einfallwinkels auf die Oberflächennormale ab. Die Beleuchtungssituation hängt daher mit der Kameraposition zusammen. Diese Reflexionen beeinflussen allerdings nur die Lichtintensität und nicht die Lichtfarbe.

Für eine farbtreue Wiedergabe ist es wichtig, die Farbtemperatur, die von der Beleuchtungsart am Aufnahmeort abhängt, zu bestimmen. Dies erfolgt über einen Weißabgleich der Kameras. Fertigungstoleranzen der Kameras können erfordern, dass der Weißabgleich bei gleichen Modellen für jede Kamera einzeln durchgeführt werden muss. Um den Weißabgleich durchzuführen, wird ein weißes Referenzobjekt mit der Kamera aufgenommen und die Gewichtung der Farben RGB so einstellt, dass die Mischfarbe zu gleichen Teilen aus den jeweiligen Grundfarben besteht. Um einer Sättigung in einem Kanal vorzubeugen, kann das Referenzobjekt auch grau sein. Allerdings ist hierbei zu beachten, dass bei sehr farbstichigem Licht der Kontrast in den einzelnen Farbkanälen stark variieren kann. Dies kann zu unterschiedlichem Rauschverhalten in den einzelnen Farbkanälen führen.

5 Homogene Koordinaten und geometrische Transformationen

Die Koordinaten werden in dieser Arbeit durchweg als homogene Koordinaten dargestellt. Dafür werden die karthesischen x/y-Bildkoordinaten um eine homogene Komponente w erweitert, die die Projektionsebene des Bildes darstellt. Ist $w = 1$, entspricht die Projektionsebene der ursprünglichen Bildebene. Für alle $w \neq 0$ gilt:

$$\begin{bmatrix} x \\ y \end{bmatrix} \rightarrow \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x \cdot w \\ y \cdot w \\ w \end{bmatrix} \tag{2}$$

Analog gilt für karthesiche Koordinaten von Punkten im 3D-Raum:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \rightarrow \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} X \cdot W \\ Y \cdot W \\ Z \cdot W \\ W \end{bmatrix} \tag{3}$$

Mit homogenen Koordinaten können auch im Unendlichen liegende Punkte über $w = 0$ dargestellt werden.

5.1 Homographien

Der Hauptvorteil homogener Koordinaten gegenüber einer Darstellung in karthesichen Koordinaten ist, dass Translationen der Punkte auch über Matrixmultiplikationen dargestellt werden können, während diese in karthesischen Koordinaten nur über eine separate Addition möglich sind:

karthesische Koordinaten

$$\begin{bmatrix} x_b \\ y_b \end{bmatrix} = \begin{bmatrix} r_1 & r_2 \\ r_3 & r_4 \end{bmatrix} \cdot \begin{bmatrix} x_a \\ y_a \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

homogene Koordinaten

$$\begin{bmatrix} x_b \\ y_b \\ 1 \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & t_1 \\ r_3 & r_4 & t_2 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_a \\ y_a \\ 1 \end{bmatrix}$$

Somit können sämtliche linearen Transformationen, unter anderem Skalierung, Scherung, Translation und Rotation, zu einer einzigen Matrixmultiplikation zusammengefasst werden. Die typischen Formen dieser Transformationen sind in Anhang A gelistet.

Für die Verarbeitung der Koordinaten werden in dieser Arbeit Homographien verwendet, die eine spezielle Teilmenge der Transformationen bilden. Eine Homographie ist eine invertierbare lineare Abbildung der realen Projektionsebene in den projektiven Raum, die Linien auf Linien abbildet. Für die Transformation eines 2D-Punktes p_a auf einen Punkt p_b mit Hilfe einer Homographie gilt:

$$\begin{gathered} p_a = \begin{bmatrix} x_a \\ y_a \\ 1 \end{bmatrix}, \; p'_b = \begin{bmatrix} x_b \cdot w \\ y_b \cdot w \\ w \end{bmatrix}, \; H_{ab} = \begin{bmatrix} h^{11} & h^{12} & h^{13} \\ h^{21} & h^{22} & h^{23} \\ h^{31} & h^{32} & h^{33} \end{bmatrix} \\ p'_b = H_{ab} \cdot p_a \\ p_b = p'_b / w = \begin{bmatrix} x_b \\ y_b \\ 1 \end{bmatrix} \end{gathered} \tag{4}$$

Bei der inversen Transformation von p_b auf p_a gilt für die Homographie:

$$H_{ba} = H_{ab}^{-1} \tag{5}$$

Dies gilt analog auch für 3D-Raumpunkte, wobei H in diesem Fall durch eine 4×4-Matrix beschrieben wird.

5.2 Projektive Transformation

Einen Sonderfall der Transformation stellt die projektive Transformation eines 3D-Punktes auf einen 2D- Punkt dar. Sie ist definiert durch eine 3×4-Projektionsmatrize P, die auch Kameramatrix genannt wird:

$$\begin{bmatrix} x \cdot w \\ y \cdot w \\ w \end{bmatrix} = P \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{6}$$

Da P singulär ist, kann man P und damit auch Gleichung 6 nicht direkt invertieren, da bei der Projektion eines 3D- Punktes auf eine 2D-(Bild)Ebene die Tiefeninformation verloren geht.

Bei der Transformation eines Raumpunktes X_a auf einen Raumpunkt X_b mit Hilfe der Homogaphie H kann eine Projektionsmatrix P_a so transformiert werden, dass die Projektionen x_a und x_b von X_a und X_b identisch bleiben:

$$\begin{gathered} x_a = P_a X_a \\ x_b = P_b X_b \\ X_b = H X_a \\ P_b = P_a H^{-1} \\ x_b = P_b X_b = P_a H^{-1} H X_a = P_a X_a = x_a \end{gathered} \tag{7}$$

Anhand von Gleichung 7 können somit Punkte und ihre Projektionen aus unterschiedlichen Koordinatensystemen in ein gemeinsames Koordinatensystem transformiert werden, wenn die dazugehörigen Homographien bekannt sind.

6 Punktmerkmale

Punktmerkmale in Bildern sind Bildpositionen, die in einer begrenzten Einflussregion gemäß ihrer (mathematischen) Modellierung möglichst einzigartig und stark ausgeprägt sind. In der Literatur werden sie auch häufig Interest-Punkte oder Feature-Punkte genannt. Die Aufgabe des mathematischen Modells ist es, dafür zu sorgen, dass ein Merkmal sich deutlich vom Hintergrund abhebt und selten genug ist, um unterscheidbar zu sein. Das Punktmerkmal ist idealerweise nicht abhängig von geometrischer - insbesondere perspektivischer - Verzerrung, Beleuchtung, Rauschen und bei bestimmten Definitionen auch Skalierung, das heißt, auch wenn die Größe des Bildes geändert wird, bleibt die Position des Punktmerkmals relativ zur Größe gleich. Zwei Arten von Bildinhalten erfüllen diese Bedingungen recht gut und werden zur Klassifizierung der Modelle benutzt: Ecken von Objektkanten (Corners) und punktförmige Objekte (Blobs).

Eine Ecke ist der Kreuzungspunkt zweier Bildkanten und kann sehr gut an Stellen lokalisiert werden, an denen zwei Bildgradienten aus unterschiedlichen Richtungen aufeinandertreffen. Die Position dieses Kreuzungspunktes ist bei idealen Linien unabhängig von der Länge die Linien. Die Größe der Einflussregion wird daher ausschließlich vom Grad der Rauschunterdrückung bestimmt. Allerdings ist aus demselben Grund die Bestimmung der Skalierung eines solchen Punktes nicht immer eindeutig, da eine Ecke über eine große Spanne von Einflussregionen immer als solche modelliert wird.

Diese Einschränkung tritt bei Blobs nicht auf: In einem gegebenen Raster gibt es genau eine Einflussregion, bei der ein kreisförmiges Objekt einem Punkt am ähnlichsten ist. Daher ist bei der Suche nach Blobs die Größe der Einflussregion sehr wichtig. Beispielsweise kann bei einem Bild eines montierten Autorads je nach Regionsgröße der Radkasten, der Reifen, die Felge, die Nabe oder sogar jede einzelne Radschraube als markant gelten. Die Größe der Einflussregion hängt daher direkt mit der Skalierung des gefundenen Blobs zusammen. Allerdings darf die Einflussregion eine bestimmte Größe nicht unterschreiten, um den Einfluss von Bildrauschen zu minimieren. Auf der anderen Seite nimmt die Positionierungsgenauigkeit relativ zur Bildgröße bei sehr großen Einflussregionen ab.

6.1 Detektoren

Drei prominente Punktmerkmalsdetektoren sind die Eckpunktdetektoren von Harris [18] und Förstner [14] sowie der Blobdetektor SIFT von Lowe[39]. Die Detektoren werden in der Literatur auch Operatoren genannt.

6.1.1 Eckdetektoren

Der Harris-Operator, auch Plessy-Punkt-Operator genannt, sowie der Förstner-Operator bestimmen Punktmerkmale über die Autokorrelationsmatrix **A** an der Stelle (i,j) der Gaußschen Richtungsableitungen f_x bzw f_y der Bildfunktion f über ein Gebiet Ω :

$$A(i,j) = \begin{bmatrix} \sum_\Omega f_x^2 & \sum_\Omega f_x f_y \\ \sum_\Omega f_x f_y & \sum_\Omega f_y^2 \end{bmatrix} \tag{8}$$

Der Harris-Operator berechnet nun eine Punktstärke an jeder Position des Bildes mit der Funktion V :

$$V = det\,(A) - k \cdot spur\left(A^2\right) \tag{9}$$

Der Parameter $k = 0.04$ wurde so gewählt, dass Punkte positive und Kanten negative Werte erhalten. Eine Non-Maxima-Unterdrückung in einem 3×3-Fenster liefert die Punktmerkmale. Vorteil dieser Technik ist, dass der einzige Parameter, der vom Nutzer bestimmt werden muss, die Größe von Ω ist.

Ein genauerer Detektor ist der Förstner-Operator [14, 56]. Die Matrix A^{-1} entspricht der Kovarianzmatrix, wodurch sich die Sicherheit in der Positionsbestimmung über Fehlerellipsen modellieren lässt. Diese Fehlerellipsen müssen eine gewisse Stärke und Rundheit aufweisen, um als Punktmerkmal zu gelten. Die Stärke w des Punktmerkmals berechnet sich über die Eigenwerte λ_1 und λ_2 der Matrix A^{-1}:

$$w = \frac{1}{\lambda_1 + \lambda_2} = \frac{det\,(A)}{spur\,(A)} \tag{10}$$

Die Rundheit q wurde in [14] ursprünglich über das normierte Verhältnis von λ_1 und λ_2 berechnet. In [59] wird jedoch beobachtet, dass, wenn λ_1 und λ_2 unterhalb der Rauschgrenze liegen und damit nicht auswertbar sind, sie wegen ihrer kleinen Zahlenwerte dennoch ein gutes Verhältnis haben können und somit irrtümlich als "gut" klassifiziert werden könnten. Da der größere der beiden Eigenwerte durch den Wertebereich der Bildpunkte nach oben beschränkt ist, ist es ausrechend, den kleineren Eigenwert gegen einen Schwellwert $thresh_{round} \epsilon\, [0.4; 2]$ zu testen.

$$\begin{gathered} q = min\,(\lambda_1, \lambda_2) = \tfrac{spur(A)}{2} - \sqrt{\left(\tfrac{spur(A)}{2}\right)^2 - det\,(A)} \\ q \geq thresh_{round} \end{gathered} \tag{11}$$

Wie beim Harris-Operator wird eine Non-Maxima-Unterdrückung auf der Stärke w durchgeführt. Danach werden sowohl die Werte für w als auch q an den verbliebenden lokalen Maxima mit ihren jeweiligen Schwellwerten verglichen und bei positiver Prüfung als gut eingestuft. Der Schwellwert für w kann relativ zum Median aller lokalen Maxima angegeben werden. Der Durchschnittswert ist nicht zu empfehlen, da die Verteilung der Punktstärke nicht gleichmäßig ist und dadurch die starken Punkte den Durchschnittswert verzerren. Der Schwellwert für q ist abhängig vom Bildrauschen. Geeignete Werte liegen zwischen 0.5 und 2. Ein Vorteil des Förstner-Operators ist, dass eine Subpixelbestimmung des Punktmerkmals möglich ist, indem man ein Paraboloid über die 8-Nachbarschaft der w-Werte mittels Kleinste-Quadrate-Ansatz approximiert und das Punktmerkmal in das Maximum des Paraboloiden schiebt.

Der Förstner-Operator kann problemlos auf Farbbilder erweitert werden. Dazu werden die Grauwertgradienten in Gleichung 8 durch die Summe der Gradienten der Farbkanäle r, g und b ausgetauscht. Der Farbkorrelationswert ergibt sich daher aus:

$$\begin{gathered} c^2 = \frac{\sum_\Omega {r_x}^2 + \sum_\Omega {g_x}^2 + \sum_\Omega {b_x}^2}{3} \\ d^2 = \frac{\sum_\Omega {r_y}^2 + \sum_\Omega {g_y}^2 + \sum_\Omega {b_y}^2}{3} \\ cd = \frac{\sum_\Omega r_x r_y + \sum_\Omega g_x g_y + \sum_\Omega b_x b_y}{3} \\ A(i,j) = \begin{bmatrix} c^2 & cd \\ cd & d^2 \end{bmatrix} \end{gathered} \tag{12}$$

In [53] ist gezeigt, dass die Gaußableitungen in der Autokorrelation zu einem systematischen Offset an Ecken führen. Dieser Offset kann durch Verwendung des Gradient-Energie-Tensors (GET) [11] anstelle der Richtungsableitung für die Autokorrelationsmatrix A minimiert werden.

6.1.2 Punktdetektoren

Ein häufig verwendeter Repräsentant der Blob-Operatoren ist der sehr universelle SIFT-Algorithmus. Hierbei wird das Bild durch Gaußfunktionen mit steigendem σ geglättet. Durch die Differenz zweier aufeinanderfolgender Glättungsbilder wird die *Difference of Gaussian* (DoG) berechnet, die eine gute Approximation der Laplaceableitung der zweidimensionalen Gaußglättung darstellt. Durch das stetige Wachsen von σ entsteht ein Stapel von DoG-Matrizen. Punktmerkmale sind lokale Maxima in der 3×3×3-Nachbarschaft in einer DoG-Matrix und ihren zwei benachbarten DoG-Matrizen. Dadurch hat das Merkmal nicht nur

eine Position, sondern auch eine Skalierung, die aus dem aktuellen Wert für σ abgeleitet werden kann. Diese Punkte sind Kandidaten für stabile Punktmerkmale und müssen noch einige Stabilitätsfilter durchlaufen. Zuerst wird anhand einer Taylor-Entwicklung der benachbarten DoG-Werte eine Subpixelposition bestimmt. Konvergiert diese innerhalb von wenigen Iterationen auf eine gültige Subpixelposition im Bereich [-0.5, 0.5], ist der Punkt stabil. Als zweites Filterkriterium wird der DoG-Wert des Kandidaten gegen einen Schwellwert geprüft. Zu kleine Werte werden wieder verworfen. Ein dritter Test filtert die Punktmerkmale, die auf langen Kanten liegen. Dazu wird eine ähnliche Funktion wie in Gleichung 9 verwendet. Für Details ist die Lektüre von [39] sehr anzuraten. Der große Vorteil des SIFT-Operators liegt in seiner Skalierungsinvarianz. Allerdings ist seine Berechnung wegen der vielen Gaußglättungen recht aufwändig und speicherintensiv.

6.2 Deskriptoren

Hat man markante Punkte im Bild lokalisiert, müssen diese so beschrieben werden, dass ein Algorithmus fähig ist, sie auch mit leichter Veränderung in anderen Bildern wiederzuerkennen. Hierbei steht man vor einem grundlegenden Problem: Wird das Merkmal mit zu wenigen Eigenschaften beschrieben, ist die Eindeutigkeit nicht gewährleistet und Fehlzuordnungen treten häufig auf. Wird das Merkmal jedoch zu differenziert beschrieben, kann es schon bei kleinsten Veränderungen, z.B. durch das Kamerarauschen, in anderen Bildern nicht wiedergefunden werden. Die Publikation der SIFT-Technik [39] hat in diesem Forschungsfeld neue Impulse gegeben, was in der unmittelbaren Folge zu vielen Varianten dieser Technik führte: Principle Component Analysis (PCA)-SIFT [32], Gradient Location and Orientation Histogram (GLOH)-Deskriptor [44], Speeded Up Robust Features (SURF)-Deskriptor [2] und Mahalanobis-SIFT (MSIFT) [43].

Jede dieser Varianten wurde entwickelt, um der ursprünglichen SIFT-Technik in bestimmten Teilbereichen überlegen zu sein, SIFT ist jedoch nach wie vor am universellsten einsetzbar und kann mit seinen Standardparametern zuverlässig sehr gute Ergebnisse liefern [43]. Aus diesem Grund wird hier nur der SIFT-Deskriptor näher beschrieben. Die anderen Techniken unterscheiden sich lediglich in manchen Details, beispielsweise durch eine vereinfachte Gaußglättung (SURF), oder benötigen Trainingsdaten, um die wichtigen Komponenten zu erkennen (MSIFT).

6.2.1 SIFT-Deskriptor

Zusätzlich zum SIFT-Merkmalsdetektor wird in [39] eine Technik beschrieben, die den Bildinhalt an der Position des Punktmerkmals beschreibt. Diese Beschreibung des Bildes generiert ein Histogramm aus den Gradientenrichtungen und -stärken. Die Histogrammwerte werden in einen rotations- und beleuchtungsinvarianten Vektor gespeichert, der 128 Elemente fasst und im folgenden Text Deskriptor genannt wird. Die Deskriptorberechnung erfolgt in zwei Abschnitten: die Bestimmung der Ausrichtung des Features und die eigentliche Histogrammberechnung mit normalisierten Richtungsinformationen.

Die Eingangsparameter des Deskriptors sind die Koordinaten des Punktmerkmals und der Skalierung des Punktes entsprechende Richtungsableitungen. Diese Richtungsableitungen werden in Polarkoordinaten umgerechnet, um ein in 10°-Schritten aufgelöstes Rotationshistogramm zu erzeugen. Dazu werden in einer skalierungsabhängigen Region alle Gradientenrichtungen mit ihrer Stärke und einer Gaußverteilung gewichtet und in ein Rotationshistogramm gespeichert. Dieses Histogramm wird nun geglättet, sein Maximum beschreibt die Hauptrichtung des Merkmalpunktes. Existieren im Histogramm weitere lokale Maxima, die relativ zum globalen Maximum einen bestimmten Schwellwert überschreiten, werden diese als weitere Hauptrichtungen gespeichert. Dies verbessert die Qualität der Beschreibung, da besonders bei Ecken mit mehreren starken Gradientenrichtungen nicht entschieden werden kann, welche Richtung die stärkere ist, und daher der Deskriptor für beide Richtungen gebildet wird.

Für jede Hauptrichtung wird nun ein richtungsnormalisiertes Histogramm der Region um das Punktmerkmal gebildet. Dazu wird die Region in 4×4 Unterregionen (Kacheln) ein-

geteilt und alle Punkte der Region auf die Hauptrichtung ausgerichtet. Die gaußgewichtete Stärke dieses Punktes wird in ein 8er-Histogramm pro Kachel eingetragen. Dies wird in Abbildung 4 anhand einer 2×2-Kachel illustriert. Der Kreis symbolisiert die Gaußgewichtung. Um Aliaseffekte zu vermeiden, wird eine trilineare Filterung zwischen benachbarten Kacheln und Histogrammunterteilungen angewendet. Das Resultat besteht aus $4 \times 4 \times 8 = 128$ Zahlenwerten, die den Deskriptorvektor bilden und euklidisch normalisiert werden. Hohe Werte des normalisierten Vektors werden abgeflacht, um zu verhindern, dass starke Strukturen schwache unterdrücken. Danach wird der Deskriptor ein weiteres Mal normalisiert.

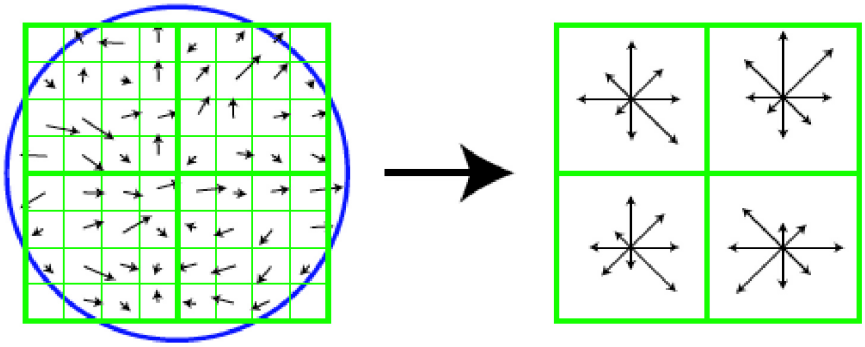


Abbildung 4: SIFT-Deskriptor

Der Deskriptor hat viele Parameter: Regionsgröße, Histogrammunterteilung für die Hauptrichtung, Gaußgewichtung, Glättung, Schwellwerte für die lokalen Maxima, Kachelgröße und Abflachung des Vektors. Zwar gibt Lowe Werte für all diese Parameter an, jedoch wird keiner dieser Werte hergeleitet oder theoretisch untermauert [43]. Dennoch scheinen die vorgeschlagenen Werte einen sehr universellen Charakter zu haben und produzieren gute Ergebnisse.

Um die Ähnlichkeit zweier Deskriptoren zu bestimmen, wird der euklidische Abstand zwischen ihnen berechnet. Lowe schlägt vor, dass das Verhältnis zwischen der besten und zweitbesten Ähnlichkeit unterhalb eines bestimmten Schwellwerts liegen sollte. So wird berücksichtigt, dass nicht korrelierbare Deskriptoren sehr hohe Abstände haben, die um Größenordnungen größer sind als eine gute Korrelation. Leider ist dieses Verfahren sehr aufwändig, da ein n-zu-n-Vergleich durchgeführt werden muss. Durch die hohe Dimensionalität des Vektors kann dieser Deskriptorenvergleich durch Abschätzung oder Vorsortierung nur unzureichend eingeschränkt werden.

7 Tracking

Videodaten können auf vielfältige Weise angewendet werden. Stationäre Kameras können Objekte verfolgen und bewegte Kameras können stationäre Objekte rekonstruieren. Die Bewegung einer Kamera kann allein aus den Bildern bestimmt werden, wenn die Kalibrierinformationen des Aufnahmesystems vorliegen [45, 54]. Eine Hauptaufgabe in der Videodatenverarbeitung besteht darin, in dichten Bildfolgen Punktmerkmale wiederzufinden (engl. tracking). Ein Korrelation des gesamten Bildes ist in der Regel nicht praktikabel, weil das Datenvolumen einer Videosequenz sehr hoch ist. Eine Videokamera mit einer Auflösung von einem Megapixel (MP) und 25 Farbbildern pro Sekunde produziert bei 8 Bit pro Farbkanal 75MB Daten pro Sekunde. Für viele Anwendungen ist es jedoch ausreichend, sich ausschließlich auf das Tracking der Punktmerkmale zu konzentrieren. Daher ist es besonders wichtig, dass die Merkmalsoperatoren echtzeitfähig sind, das heißt, die Zeitdifferenz zwischen Eingabe und Antwort darf eine gewisse Grenze nicht überschreiten.

Die einzelnen Bilder unterscheiden sich in der Regel nur minimal. Daher können bestimmte Annahmen gemacht werden, ohne das System zu stark einzuschränken. Wegen der dichten Bildfolge kann der Suchradius für ein Punktmerkmal stark verkleinert werden. Die perspektivische Verzerrung und die Skalierung der einzelnen Objekte zweier aufeinanderfolgender

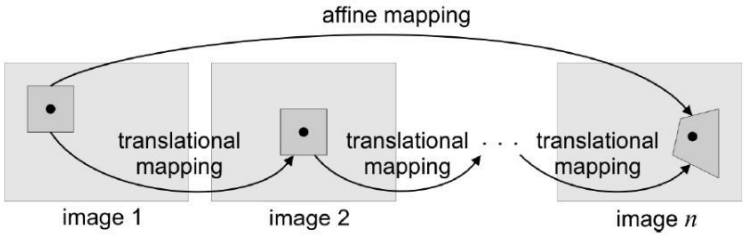


Abbildung 5: Affine Konsistenzprüfung beim Standard-KLT

Bilder variiert in der Regel ebenfalls ist nicht sehr stark, wodurch die Suche nach korrespondierenden Punktmerkmalen vereinfacht wird. Die grundlegenden Arbeiten von [59, 40] umfassen viele Aspekte des Trackings von Videodaten und werden hier exemplarisch vorgestellt.

7.1 Standard KLT

Zwei weit verbreitete und frei verfügbare Implementierungen des KLT Feature Tracker sind von Birchfield und [5] (Weblinks siehe Anhang B). Der KLT Feature Tracker berechnet den optischen Fluss der Punktmerkmale mittels Bildpyramiden. Um Ausreißer zu vermeiden, werden die jeweils aktuellen Punktmerkmale mit einem Referenzmerkmal verglichen, das der ersten Beschreibung des erfolgreich verfolgten Punktes entspricht. Daher addieren sich die perspektivischen Fehler zum Referenzmerkmal von Bild zu Bild auf (vgl. Abbildung 5).

Die Punktmerkmale werden bisher in allen Implementierungen nur auf Grauwertbildern berechnet. Die Implementierung von [5] ist sehr effizient und bestimmt die Punktmerkmale mit Subpixelgenauigkeit, benutzt jedoch statt der genaueren Gauß ´schen Richtungsableitungen einfache Intensitätsdifferenzen als Bildableitungen. Die Implementierung von Birchfield umfasst eine affine Konsistenzprüfung, um die Fehlerakkumulation durch perspektivische Verzerrungen zu begrenzen und verwendet Gauß ´sche Richtungsableitungen. Daher ist diese Implementierung nicht so schnell und weniger für Echtzeitanwendungen geeignet. Eine für Grafikprozessoren (GPU) angepasste Version des KLT Feature Tracker verspricht Echtzeitfähigkeit bis zu einer Auflösung von 1024×768 Bildpunkten bei 30Hz [60].

8 Räumliches Stereo

Räumliches Stereo bezeichnet die Berechnung von 3D-Objektpunkten mittels Triangulation. Dazu wird ein Objekt von mindestens zwei verschiedenen Standpunkten aus fotografiert. Ein Bild wird Referenzbild und das andere Sekundärbild genannt. Analog würde ein eventuell vorhandenes drittes Bild Tertiärbild genannt werden. Als zugrunde liegendes Kameramodell wird die ideale Lochkamera angenommen. Bei diesem Modell fallen alle einfallenden Lichtstrahlen durch ein Loch, das Projektionszentrum, auf eine photosensitive Projektionsfläche, die Bildebene, und projizieren das Bild der Szene. Zwei Lichtstrahlen, die vom selben Objekt ausgehen und in beiden Kameras abgebildet werden, bilden zusammen mit der Verbindungslinie der Bildpunkte auf der Bildebene ein Dreieck (vgl. Abbildung 1). Diese Bildpunkte aus unterschiedlichen Bildern, die dasselbe Objekt zeigen, sind so genannte Bildkorrespondenzen. Sind alle bestimmbaren, gültigen Bildkorrespondenzen in zwei Bildern gefunden, gelten die Bilder als registriert.

Mit Hilfe dieser Bildkorrespondenzen und der Position der Kameras kann die 3D-Position des Objektes bestimmt werden. Während die Kamerapositionen nur einmal bestimmt werden müssen, muss die Suche nach Bildkorrespondenzen für jeden Bildpunkt einzeln durchgeführt werden. Da Schwierigkeiten wie perspektivische Verzerrungen, Verdeckungen, homogene Flächen und Mehrdeutigkeiten auftreten können und das Finden von Bildkorrespondenzen durch die Verdeckungen eine nicht vollständig lösbare Aufgabe ist, wird die Suche nach Bildkorrespondenzen als "Korrespondenzproblem" bezeichnet. Allerdings kann man das Problem

etwas entschärfen. Die Bildkorrespondenzen können sich nicht an beliebigen Stellen im Bild befinden, sondern müssen auf berechenbaren Geraden, den so genannten Epipolargeraden, liegen. Sind alle Epipolargeraden parallel zu den Zeilen des Bildrasters, spricht man vom Stereonormalfall: Die Bildkorrespondenz im Sekundärbild muss auf der gleichen horizontalen Linie liegen wie im Referenzbild. Der Abstand der Punkte auf dieser Linie wird Disparität genannt. Die Berechnung der Epipolargeraden, die genauen Eigenschaften von Stereonormalbildern und ihre Erzeugung werden in Abschnitt 9.4 erläutert. Im weiteren Text wird davon ausgegangen, dass die Bildzeilen mit den Epipolargeraden übereinstimmen.

Die Position einer Bildkorrespondenz auf der Epipolargeraden hängt mit dem Abstand des Objektes zur Kamera zusammen. Jeder Punkt im Referenzbild erzeugt somit eine Gerade durch das Sekundärbild, deren Korrespondenzposition von der Tiefe des Objektes abhängt. Da die Epipolargerade den Referenzpunkt schneidet, kann die Position der Korrespondenz durch den Abstand zur Referenz im Folgenden als Disparität bezeichnet werden. Die Werte werden für jeden Punkt in einer Disparitätenmatrix D gespeichert und es gilt für alle zugeordneten Bildpunkte:

$$I_1(x, y) = I_2(x + D(x, y), y) \tag{13}$$

Es können kleinst- und größtmögliche Disparitäten angegeben werden, die durch das Pixelraster auf ganzzalige Werte begrenzt sind. Verknüpft man diesen linearen Disparitätsbereich mit der Position im Referenzbild, entsteht ein 3D-Suchraum, in dem das reale Objekt gefunden werden muss.

Durch die genannten Einschränkungen ist es im Allgemeinen nicht möglich, für jeden Bildpunkt eine eindeutige Korrespondenz zu finden. Es bedarf daher robuster Lösungsansätze für das Korrespondenzproblem, die Entwicklung dieser Ansätze ist ein Schwerpunktthema der digitalen Bildverarbeitung. Auf der Website der Middleburry University [58] können auf einem standardisierten Testdatensatz verschiedene Verfahren miteinander verglichen werden. Die meisten Verfahren versuchen die geschätzte Lösung mittels einer Energie- oder Kostenfunktion zu modellieren. Dieser Ansatz basiert auf der Annahme, dass alle Korrespondenzen in dieser Kostenfunktion einen kleinen Kostenanteil verursachen und die richtige globale Lösung in der Summe die global niedrigsten Kosten erzeugt. Die meisten Funktionen bestehen aus zwei Teilen: einer lokalen Ähnlichkeitsfunktion und einer globalen Glättungsfunktion, die versucht, durch zusätzliche Annahmen wenig plausible Tiefensprünge zu verhindern oder bei ähnlichen lokalen Kosten die wahrscheinlich richtige Korrespondenz auszuwählen. Vertretbare Annahmen sind u. a., dass die Szene stückweise glatt ist und die Tiefenreihenfolge erhalten bleibt.

Die Güte einer Methode ist also in der Modellierung ihrer Kostenfunktion verankert. Problematisch ist allerdings die Bestimmung des globalen Optimums, da die Kosten eines Punktes abhängig von den Nachbarpunkten sind. Theoretisch müsste dazu jede Permutation aller Korrespondenzmöglichkeiten gebildet und die entsprechende Gesamtenergie dieser Permutation berechnet werden, da eine analytische Bestimmung des Minimums in der Regel nicht möglich ist. Die Anzahl dieser Permutationen steigt exponentiell mit der Bildgröße und die Berechnung sämtlicher Kombinationsmöglichkeiten ist daher schon bei kleinen Bildern nicht praxistauglich. Deshalb muss für die Entwicklung eines effizienten Algorithmus, das heißt eines Algorithmus mit polynomineller Laufzeit, in Kauf genommen werden, das globale Minimum nur zu approximieren. Hierbei wirkt sich positiv aus, dass entfernte Bildbereiche sich nur sehr selten direkt beeinflussen und daher eine Begrenzung des Nachbarschaftseinflusses naheliegend ist. Der aktuelle Stand der Forschung für diese Approximation ist:

- Dynamic Programing (DP): Eine der ersten Optimierungsmethoden für das Korrespondenzproblem ist die Dynamische Programmierung [3] entlang von Epipolarlinien. Hierbei wird davon ausgegangen, dass die beste Lösung für das Teilproblem entlang dieser Epipolarlinie zum globalen Optimum führt. Diese Strategie gehört damit zu den Greedy-Algorithmen, die immer erfolgreich sind, wenn die Kostenfunktion keine lokalen Extrema hat [7]. Die Qualität dieser Technik ist mit aktuellen Verfahren konkurenzfähig, wenn vertikale Konsistenz über mehrere Epipolarlinien sichergestellt wird [36].

- Belief propagation (BP): Diese Methode modelliert das Korrespondenzproblem als Markov-Feld und versucht, über eine "Bayesian belief propagation" das a-posteriori-Optimum zu schätzen.[34].
- Graph Cuts (GC): Dieses Verfahren basiert auf der Graphentheorie und versucht den Graphen anhand des maximalen Flusses durch einen Knoten zu unterteilen [35]. Die Aufgabe besteht nun darin, die Energiefunktion so als gerichteten Graphen zu formulieren, dass ein minimaler Schnitt des Graphen auch die Energiefunktion minimiert.
- Semi-Global Matching (SGM): Dieses relativ neue Verfahren versucht den Suchraum durch ein relativ einfaches Gradientenverfahren einzugrenzen [27, 26]. Auch dieses Verfahren gehört in die Klasse der Greedy-Algorithmen und findet daher in relativ kurzer Zeit garantiert ein lokales Optimum. Die Gradientenanstiege werden aus mindestens acht Richtungen durch den Suchraum berechnet und aufsummiert. Dabei gibt es nur zwei zusätzliche Kostenfaktoren: einen für kleine Disparitätsänderungen und einen für Tiefensprünge. Daher können selbst leicht fluchtende Ebenen auch über große Bildbereiche gut modelliert werden. Das Verfahren wurde ursprünglich für den Luftbildfall entwickelt, ist aber universell einsetzbar und besticht durch seine geringe Anzahl von Parametern.

8.1 Lokale Ähnlichkeitsfunktion

Eine lokale Ähnlichkeitsfunktion $\varrho(I_1, I_2)$ berechnet die Ähnlichkeit für zwei Bildpositionen in den Bildern I_1 und I_2 der Größe $M = w \times h$. Die Funktionen können in zwei Kategorien unterteilt werden: in fensterbasierte und punktbasierte Funktionen. Bei RGB-Farbbildern werden die einzelnen Kanäle gemeinsam betrachtet, wodurch die Anzahl der betrachteten Messungen verdreifacht wird. Die fensterbasierten Ähnlichkeitsfunktionen berechnen sich über zwei $n \times n$ Punkte große Bildausschnitte $a(x_1, y_1)$ und $b(x_2, y_2)$. Um die Notation zu vereinfachen, wird im Folgenden die Verschiebung (x_1, y_1) und (x_2, y_2) weggelassen und es wird nur die relative Position zu dieser Stelle im Bild angegeben. Gebräuchliche Funktionen sind die Summe der quadratischen Abstände (engl. *sum of squared differences - SSD*):

$$\varrho_{SSD}(a,b) = \frac{1}{3 \cdot M} \sum_{i,j=1}^{n,m} \sum_{k=R,G,B} \left(a(i,j) - b(i,j)\right)^2 \tag{14}$$

und die geringfügig schneller zu berechnende Summe der absoluten Abstände (engl. *sum of absolute differences - SAD*):

$$\varrho_{SAD}(a,b) = \frac{1}{3 \cdot M} \sum_{i,j=1}^{n,m} \sum_{k=R,G,B} |a(i,j) - b(i,j)| \tag{15}$$

Bei diesen Funktionen können die Farbkanäle auf dieselbe Intensität normalisiert werden, um Helligkeits- und Kontrastunterschiede zu kompensieren. Das zu verwendende intensitätsnormalisierte Bild berechnet sich aus:

$$I' = \frac{I}{R + G + B} \tag{16}$$

Diese Messwerte müssen nun noch auf das Intervall [0;1] normalisiert werden, wobei der Wert 1 für die maximale Ähnlichkeit steht:

$$\varrho'(a,b) = \frac{1}{1 + \varrho(a,b)} \tag{17}$$

Sehr häufig wird die normierte Kreuzkorrelation (engl. *normalized cross-correlation - NCC*) verwendet, da sie helligkeits- und kontrastinvariant ist.

$$\begin{gathered}\varrho_{NCC}(a,b) = \frac{\sigma_{ab}}{\sqrt{\sigma_a^2 \cdot \sigma_b^2}} \\ \sigma_{ab} = \frac{1}{3 \cdot M}\left(\sum_{i,j=1}^{n,m}\sum_{k=R,G,B} a(i,j) \cdot b(i,j)\right) - \overline{a} \cdot \overline{b} \\ \sigma_a = \frac{1}{3 \cdot M}\left(\sum_{i,j=1}^{n,m}\sum_{k=R,G,B} a(i,j)^2\right) - \overline{a} \\ \overline{a} = \frac{1}{3 \cdot M}\sum_{i,j=1}^{n,m}\sum_{k=R,G,B} a(i,j)\end{gathered} \tag{18}$$

Die Berechnungsgeschwindigkeit der NCC kann signifikant erhöht werden, wenn die für alle Bildkorrespondenzen konstanten Werte für σ_a, σ_b, $\overline{a}$ und $\overline{b}$ einmal vorberechnet werden. Allerdings bestimmt die NCC ausschließlich Intensitätsänderungen, weshalb homogene Flächen nicht berechnet werden können. Der Wert unter der Wurzel und damit der Quotient aus Gleichung 18 ergibt bei einer homogenen Fläche immer Null. Daher wird in [10] eine Modifikation (*MNCC*) vorgeschlagen, anhand derer zumindest eine homogene Fläche mit einer strukturierten verglichen werden kann.

$$\varrho_{MNCC}(a,b) = \frac{2\sigma_{ab}}{\sigma_a^2 + \sigma_b^2} \tag{19}$$

Beide Kreuzkorrelationsverfahren liefern Werte im Intervall [-1, 1]. Um im selben Intervall wie SSD und SAD zu liegen, müssen die Werte leicht angepasst werden:

$$\varrho'(a,b) = \frac{1 + \varrho(a,b)}{2} \tag{20}$$

Punktbasierte Funktionen berechnen die Ähnlichkeit ausschließlich über die zwei Bildpunkte an den Bildpositionen. Daher sind sie sehr rauschempfindlich und die Bilder müssen vorher in Farbe, Helligkeit und Kontrast angepasst sein. Der Ansatz in Formel 14 und 15, die Fenstergröße auf Eins zu setzten, ist nicht empfehlenswert, weil besonders an Kanten starke Aliaseffekte auftreten. Daher wird in [4] vorgeschlagen, entlang der Epipolarlinie gewichtet zu interpolieren:

$$\begin{aligned}\varrho_{BT}(x,y) = \quad & \frac{1}{3}\sum_{k=R,G,B}\left(\frac{|a(x,y)-b(x-1,y)|}{4} + \right. \\ & \left.\frac{|a(x,y)-b(x,y)|}{2} + \frac{|a(x,y)-b(x+1,y)|}{4}\right)\end{aligned} \tag{21}$$

In der Praxis zeigt sich, dass durch unterschiedlichen Weißabgleich, Rauschen und Produktionsschwankungen beim Kamerasensor die Farbtreue nicht vollständig gewährleistet ist. Alle Verfahren setzten eine gleiche Farbverteilung oder lineare Farbabstände voraus. Bilder mit unterschiedlichen Gammakurven oder von unterschiedlichen Sensoren können so nicht zuverlässig bearbeitet werden.

Ein auf der Informationstheorie basierendes Verfahren ist die Berechnung des Transinformationsgehaltes (engl. *mutual information - MI*). Die MI interpretiert beide Bildfunktionen eines Farbkanals als Zufallsvariablen und versucht den statistischen Zusammenhang dieser Variablen zu bestimmen. In dieser Arbeit wird nur auf die Berechnung einer punktweisen MI nach [27] eingegangen. Eine gute grundlegende Beschreibung der MI wird in [42, 33, 51] gegeben, ein Vergleich mit der MNCC ist in [10] aufgeführt.

Zunächst wird die Wahrscheinlichkeit eines Farbwertes i in einem Bild berechnet, was einer Histogrammberechnung aller registrierten Bildpunkte M_{reg} entspricht:

$$P_1(i) = \frac{1}{M_{reg}}\sum^{x_1} T\left[I_{1x_1} = i\right] \tag{22}$$

wobei T eine boolesche Funktion ist, die den Wert 1 liefert, wenn die Bedingung in den eckigen Klammern wahr ist, und 0 in allen anderen Fällen. Zur Berechnung der MI werden jetzt noch die gemischten Wahrscheinlichkeiten benötigt, also die Wahrscheinlichkeit des Farbwerts i an der Position $x_1$des Bildes I_1 unter der Voraussetzung, dass an der korrespondierenden Position $x_2$des Bildes I_2 der Farbwert j ist :

$$P_{1,2}(i,j) = \frac{1}{M_{reg}}\sum^{x_1 x_2} T\left[I_{1x_1} = i \wedge I_{2x_2} = j\right] \tag{23}$$

Die Bestimmung der lokalen Ähnlichkeit der MI klingt nach einem Henne-Ei-Problem, da zur Berechnung der Kosten von Bildkorrespondenzen genau diese benötigt werden. Allerdings ist die Wahrscheinlichkeitsverteilung der Gleichung 23 theoretisch nicht abhängig von der Bildgröße und bei genügend großen Bildern zeigt sich in der Praxis, dass die MI über mehrere Skalen sehr konstant ist. Daher kann die MI abgeschätzt werden, wenn die Bilder zunächst verkleinert und dann hierarchisch registriert werden, wobei in der kleinsten Auflösung ein anderes Verfahren zur Berechnung der lokalen Ähnlichkeit verwendet wird, z.B. die MNCC.

Aus diesen Wahrscheinlichkeiten kann die relative Entropie h der Farbwerte i und j berechnet werden:

$$\begin{aligned} h_{I_1}(i) &= \frac{1}{M_{reg}} log\left(P_1\left(i\right) \otimes g\left(i\right)\right) \otimes g\left(i\right) \\ h_{I_1,I_2}(i,j) &= \frac{1}{M_{reg}} log\left(P_{1,2}\left(i,j\right) \otimes g\left(i,j\right)\right) \otimes g\left(i,j\right) \end{aligned} \tag{24}$$

Die Faltung $\otimes g\left(\right)$ mit einer Gaußfunktion ist notwendig, um mittels einer Parzen-Fenster-Schätzung [33] Diskretisierungseffekte zu vermeiden. Das σ dieser Gaußfaltung ist vom Rauschverhalten des Sensors abhängig. Des weiteren muss eine minimale Wahrscheinlichkeit festgelegt werden, um zu vermeiden, dass das Argument des Logarithmus Null ist. Ein guter Wert hierfür ist die Hälfte der niedrigsten gemessenen Wahrscheinlichkeit. Die punktweise MI ist nun definiert durch:

$$mi_{I_1,I_2}\left(i,j\right) = h_{I_1}\left(i\right) + h_{I_2}\left(j\right) - h_{I_1,I_2}\left(i,j\right) \tag{25}$$

Dies führt für jeden Farbkanal zu einer 256×256-Matrix, die als Lookuptabelle des Ähnlichkeitsmaßes dient. Die Normierung dieser Lookuptabelle erfolgt über die Extremwerte:

$$\varrho_{MI}\left(i,j\right) = \frac{mi_{I_1,I_2}\left(i,j\right) - mi_{min}}{mi_{max} - mi_{min}} \tag{26}$$

8.2 Symmetrisches Stereo

Die Verfahren der Korrespondenzsuche sind leider nicht frei von Fehlzuordnungen. Um viele dieser Ausreißer detektieren zu können, wird häufig die Technik des symmetrischen Stereos angewandt. Dabei muss jeder Punkt aus dem Primärbild eineindeutig mit einem Punkt aus dem Sekundärbild verbunden werden, oder aber er hat gar keine Korrespondenz. Hierzu werden Referenz- und Sekundärbild vertauscht und die Disparitätenkarte D_{sym} für dieses Paar wird neu berechnet. Alle gültigen Zuordnungen müssen jeweils aufeinander zeigen:

$$D_{sym}\left(x + D\left(x,y\right), y\right) = -D\left(x,y\right)) \tag{27}$$

Da durch das Bildraster eine gewisse Toleranz nötig ist, um nicht zu viele gültige Korrespondenzen zu verwerfen, wird Gleichung 27 gegen einen Schwellwert getestet, der einer Rasterbreite entspricht:

$$\left|D_{sym}\left(x + D\left(x,y\right), y\right) + D\left(x,y\right)\right| \leq 1 \tag{28}$$

Die meisten Fehlzuordnungen, die zum Beispiel durch Verdeckungen entstehen, erfüllen dieses Kriterium nicht und können somit gefiltert werden. Allerdings verdoppelt sich dadurch der Berechnungsaufwand und das gefilterte Ergebnis ist nicht garantiert frei von Ausreißern.

8.3 Subpixelberechnung

Durch die Diskretisierung des Pixelrasters können an Kanten, die nicht exakt auf dem Raster liegen, sogenannte Alias-Effekte auftreten, wodurch sich die Farben von benachbarten Pixeln vermischen. Da dieser Effekt durch die Gewichtung der umliegenden Pixel verursacht wird, kann man die Position im Subpixelbereich über die benachbarten Bildpunkte approximieren. Dazu wird ein kleines Fenster im Referenzbild mit Fenstern um die entlang der Epipolargeraden ligenden Punkte mit der ganzzahligen Korrespondenz verglichen. Ist der Vergleichswert der ganzzahligen Korrespondenz s_0 höher als ihre beiden Nachbarn s_{-1} und s_{+1}, wird versucht, eine Parabel durch die drei Punkte zu legen, deren Maximum die

gewünschte Subpixelapproximierung angibt. Die Vergleichsfunktion muss kubisch sein [26], wie es beispielsweise auf die SSD zutrifft.

$$s_i = \varrho_{SSD}\left(a\left(x,y\right), b\left(x + D\left(x,y\right) + i\right), y\right) \tag{29}$$

Die Subpixelinformation lässt sich nun durch eine einfache Extremwertsuche bestimmen. Da die Abstände der Pixel zueinander gleich sind, kann die Extremwertrechnung zu folgender Gleichung vereinfacht werden:

$$D_{sub}\left(x,y\right) = D\left(x,y\right) + \frac{s_{-1} - s_{+1}}{2\left(s_{+1} + s_{-1} - 2s_0\right)} \tag{30}$$

Allerdings führt bereits leichtes Rauschen zu erheblichen Fehlern in der Approximation. Zusätzlich können homogene Flächen ohne lokales Extremum auf diese Weise nicht approximiert werden. In [15] wird daher vorgeschlagen, eine Glättung einzuführen, die sensitiv bezüglich der räumlichen Glätte und der Stärke der Subpixelbestimmung ist. Da sich die beiden Komponenten gegenseitig beeinflussen, wird eine Energiefunktion minimiert, indem sie iterativ berechnet wird, bis die Änderungen der Subpixelapproximation marginal ist.

9 Kamerakalibrierung

Um den Bezug zwischen homogenen Bildpunkten x und Raumpunkten X herzustellen, wird die Kamera durch eine Lochkamera modelliert und durch eine (3×4)-Projektionsmatrix P parametrisiert. Diese Projektionsmatrix bildet nun die 3D-Raumpunkte auf die Bildebene ab:

$$x = PX \tag{31}$$

Dabei ist zu beachten, dass P nicht eindeutig ist. Das Lochkameramodell erlaubt eine Zerlegung von P in eine innere Kalibrierung K, eine Rotation R und einen Translationsvektor t. Wobei K und R (3×3)-Matrizen sind und t ein 3-Vektor ist:

$$P = K\left[R\,|t\right] \tag{32}$$

Dabei wird die Bestimmung von K intrinsische Kalibrierung und die Bestimmung von R und t Orientierung genannt. Beides zusammen ergibt die Kamerakalibrierung. Zusätzlich wird oft das Projektionszentrum C benötigt; dieses ist definiert durch:

$$\begin{array}{c} P = KR\left[I\,|{-C}\right] \\ \Rightarrow \\ C = -R^T t \end{array} \tag{33}$$

9.1 Intrinsische Kalibrierung

Die intrinsische Kalibrierung beschreibt ausschließlich die Kameramatrix K einer Lochkamera. Das Modell der Lochkamera ist geradentreu, das heißt, Geraden werden linear im Bild abgebildet. Reale Linsen weisen zusätzlich nichtlineare Verzeichnungen auf, daher wird im Abschnitt 9.3 beschrieben, wie diese Verzerrungen korrigiert werden können. Die Matrix K ist eine obere Dreiecksmatrix mit fünf Freiheitsgraden der Form:

$$K = \begin{bmatrix} f_x & s & d_x \\ 0 & f_y & d_y \\ 0 & 0 & 1 \end{bmatrix} \tag{34}$$

Diese fünf intrinsischen Parameter beschreiben die Kamerakonstanten f_x und $f_y$in x- und y-Richtung, die Position des Bildhauptpunktes (d_x, d_y)und eines Bildscheerungsparameter s. Ein vereinfachtes Kameramodell berücksichtigt, dass die Scheerung bei den meisten Kameras Null ist und die Kamerakonstanten gleich sind:

$$K_{simple} = \begin{bmatrix} f & 0 & d_x \\ 0 & f & d_y \\ 0 & 0 & 1 \end{bmatrix} \tag{35}$$

Zu beachten ist, dass normale Linsen, auch Festbrennweiten, bei der Fokussierung und bei Änderung der Blendenöffnung die Kamerakonstante und den Bildhauptpunkt leicht verändern können. Daher muss für eine exakte Rekonstruktion die innere Kalibirierung K für die bei der Aufnahme verwendete Linseneinstellung bestimmt werden.

9.2 Orientierung

Für die Prozessierung von Stereobildern ist es unabdingbar, die beteiligten Kameras zu orientieren. Hierbei wird zwischen zwei verschiedenen Arten der Orientierung unterschieden: projektive Orientierung und kalibrierte Orientierung. Die kalibrierte Orientierung wiederum wird in absolute und relative Orientierung unterteilt.

Bei der projektiven Orientierung wird ausschließlich ein Modell gesucht, das die Bildgeometrie der Bilder zueinander beschreibt. Im Zweibildfall wird die Orientierung über die Fundamentalmatrix (Abschnitt 9.4) modelliert, im Dreibildfall über den Trifokaltensor (Abschnitt 9.5). Die Berechnung beider Modelle ist nicht eindeutig und die tatsächlichen Kamerapositionen können im Allgemeinen nicht rekonstruiert werden.

Ist die intrinsische Kalibrierung der Kameras bekannt, kann eine relative Orientierung aller Kameras zu einer Referenzkamera bis auf einen globalen Skalierungsfaktor über die Elementarmatrix (Abschnitt 9.6) berechnet werden. Die Skalierung kann über ein größengenormtes Objekt im Bild oder einen bekannten Abstand zweier Kameras bestimmt bzw. angegeben werden. Sind absolute Position und Ausrichtung der Referenzkamera bezüglich eines globalen Koordinatensystems bekannt, können die relativ orientierten Kameras in absolut orientierte Kameras transformiert werden.

Eine weitere Möglichkeit zur absoluten Orientierung besteht darin, bekannte Referenzpunkte im Raum zu identifizieren und so einen räumlichen Rückwärtsschnitt und damit die vollständige Kamerakalibrierung zu berechnen. Da es sich dabei um eine Standardtechnik handelt, wird hier darauf nicht weiter eingegangen. Genauere Informationen zum räumlichen Rückwärtsschnitt können in [16] nachgelesen werden.

Die einfachste Kameramatrix ist die kanonische Kamera: Sie befindet sich im Ursprung, hat keine Rotation, der Bildhauptpunkt liegt ebenfalls im Ursprung und die Kamerakonstante ist Eins:

$$P_K = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{36}$$

9.3 Geradentreue Bilder

Werden Geraden im Objektraum auch auf Geraden im Bild projiziert, spricht man von geradentreuen Bildern. Eine ideale Lochkamera liefert solche Bilder, reale Kameraobjektive bestehen jedoch aus mehreren Linsen, die unterschiedliche Verzeichnungen im Bild verursachen. Diese Abbildungsfehler - auch Aberrationen genannt - können kompensiert werden, wenn die ideale Projektionsposition des Objektpunktes und eine Abbildungsvorschrift zur realen Projektionsposition bekannt sind. Es ist jedoch messtechnisch nicht praktikabel, diese Abbildungsvorschrift unabhängig für jeden Bildpunkt zu bestimmen. Da der Abbildungsfehler meist aus den Fertigungseigenschaften der Linse resultiert, können manche Fehler physikalisch modelliert werden. Der wichtigste Abbildungsfehler ist die radiale Verzeichnung. Hierbei wird eine radial symmetrische Verzeichnung angenommen, die abhängig vom Abstand zu einem Symmetriepunkt ist. Der Symmetriepunkt befindet sich in der Regel nahe der Bildmitte. Da die meisten Objektive mehr als eine Linse enthalten, überlagern sich die einzelnen Verzeichnungen und es müssen komplexere Verzeichnungsmodelle gewählt werden. Das in dieser Arbeit verwendete Modell besteht aus drei radialen Verzeichnungsparametern

κ_1, κ_2 und κ_3 sowie dem Symmetriepunkt $s = (s_x, s_y)$. Damit kann aus der geradentreuen Position x in einem Bild I und dem Abstand r von Position x zum Symmetriepunkt s die verzeichnete Position $\hat{x}$ im Bild $\hat{I}$ berechnet werden:

$$\begin{aligned} r &= (x_x - s_y)^2 + (x_x - s_y)^2 \\ \hat{x} &= s + \left(1 + \kappa_1 r + \kappa_2 \left(r^2\right) + \kappa_3 \left(r^3\right)\right) x \end{aligned} \tag{37}$$

Bei bekannten Parametern kann das verzeichnete Bild aus dem idealen Bild abgeleitet werden. Die direkte Invertierung der Verzeichnungsfunktion 37 ist nicht trivial, weshalb die Entzerrung des Bildes I indirekt durchgeführt wird: Die unbekannte entzerrte Bildgröße wird über die Größe des verzerrten Bildes geschätzt oder manuell festgelegt. Nun wird für jeden Bildpunkt x berechnet, von welcher Position $\widetilde{x}$ im verzeichneten Bild er abstammt. Da diese Position sich bei fixierten Kameraeinstellungen nicht ändert, kann für die Entzerrung aller Bilder einer Kamera eine Transformationsmatrix generiert werden, die für jede Position x des entzerrten Bildes die ursprüngliche Position $\widetilde{x}$ im verzerrten Bild speichert. Die Position $\widetilde{x}$ liegt in der Regel nicht auf dem Punkteraster des verzerrten Bildes. Daher sollte der Farbwert für diese Zwischenposition nach Gleichung 109 interpoliert werden.

9.4 Projektiver Zweibildfall: Epipolargeometrie und Fundamentalmatrix

Die Theorie der Epipolargeometrie geht auf die Anfänge der Photogrammetrie zurück [63] und entwickelte sich mit der digitalen Bildverarbeitung zu dem Grundbaustein der Bildorientierung [37]. Das grundlegende Prinzip ist in Abbildung 6 zu sehen: Alle Objektpunkte entlang eines Sichtstrahls werden auf denselben Punkt x der Bildebene I_1 der Referenzkamera abgebildet. Projiziert man diesen Sichtstrahl auf die Bildebene I_2 der zweiten Kamera, erhält man die sogenannte Epipolarlinie oder Epipolargerade l des Bildpunktes x. Die Bildkorrespondenz x' des Objektpunktes X muss sich auf dieser Geraden befinden.

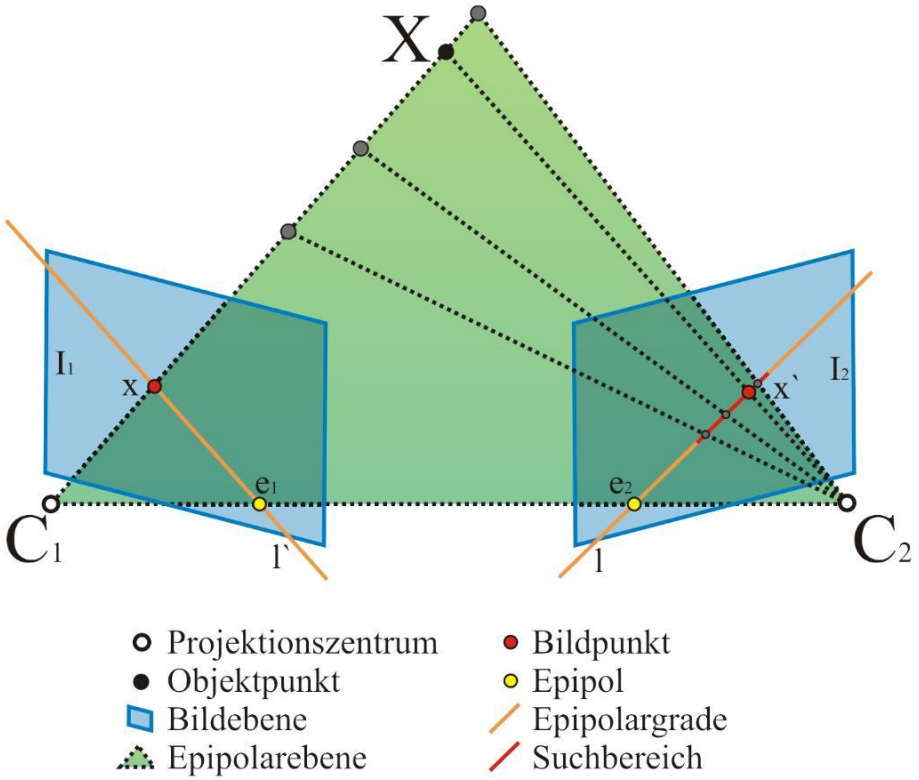


Abbildung 6: Epipolargeometrie

Die Ebene, die durch die Projektionszentren C_1, C_2 und den Objektpunkt X aufgespannt wird, ist die Epipolarebene, die Schnitte dieser Ebene mit den Bildebenen sind die Epipolarlinien. Da alle abgebildeten Sichtstrahlen das Projektionszentrum schneiden, haben alle Epipolargeraden eines Bildes einen gemeinsamen Schnittpunkt. Dieser Punkt wird Epipol genannt und ist die Projektion des Projektionszentrums C_1auf die Bildebene I_2 und umgekehrt.

Die Bestimmung der Epipolargeraden l und l' erfolgt über die 3×3-Fundamentalmatrix F. Die Parametrisierung der Gerade in homogenen Koordinaten erfolgt durch die Geradengleichung:

$$l^T x = 0 \tag{38}$$

Für die Fundamentalmatrix gelten folgende Gleichungen:

$$\begin{array}{ll} (1) & Fx = l \\ (2) & F^T x' = l' \\ (3) & x'^T F x = 0 \\ (4) & Fe_1 = 0 \\ (5) & F^T e_2 = 0 \end{array} \tag{39}$$

Die Fundamentalmatrix hat einen Rang von Zwei, daher gilt für die Determinante:

$$det\,(F) = 0 \tag{40}$$

Die Fundamentalmatrix hat sieben Freiheitsgrade und lässt sich entweder direkt aus den Projektionsmatrizen ableiten oder bei sieben Bildkorrespondenzen mit dem minimalen 7-Punktealgorithmus [37, 68] sowie bei mehr Bildkorrespondenzen mit dem normalisierten 8-Punktealgorithmus [19]bestimmen. Die 7- und 8-Punktealgorithmen bedingen, dass sich nicht alle Bildkorrespondenzen auf derselben Ebene im Raum befinden.

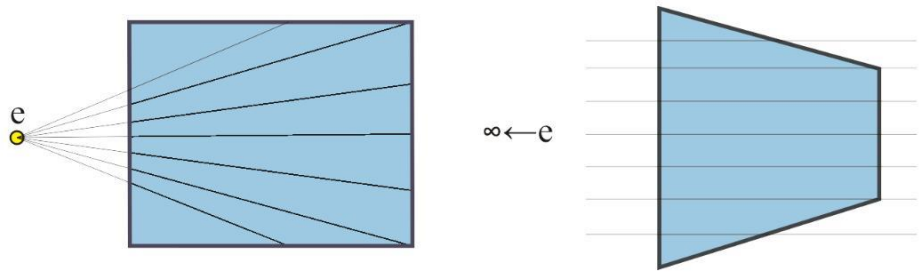


Abbildung 7: Rektifizierung zum Stereonormalfall

Liegen die Bildebenen parallel zueinander, kann das Projektionszentrum nicht auf die Bildebene abgebildet werden und der Epipol liegt im Unendlichen. Infolgedessen sind alle Epipolargeraden parallel zueinander. Liegt der Epipol auf dem Horizont, spricht man vom Stereonormalfall und die Epipolargeraden stimmen mit den Bildzeilen überein. Dies ist für die automatische Bildanalyse von Vorteil, weil das Bild Zeile für Zeile abgearbeitet werden kann. In der projektiven Geometrie kann durch die sogenannte Rektifizierung jedes Bild so verzerrt werden, dass die neue virtuelle Bildebene den Kriterien des Stereonormalfalls entspricht, indem der Epipol ins Unendliche transferiert wird (vgl. Abbildung 7). Ein gängiges Verfahren zur linearen Rektifizierung wird in |21| erklärt. Hierbei ist jedoch zu beachten, dass der Epipol nicht im Bild liegen sollte, weil sonst die Bilder ebenfalls "in die Unendlichkeit" verzerrt werden. Liegt der Epipol im Bild, kann das Bild unter bestimmten Voraussetzungen dennoch entzerrt werden [49]. Dieses Verfahren wird jedoch hier nicht verwendet.

Da im Stereonormalfall sowohl x als auch x' auf derselben Bildzeile liegen, ist der einzige tiefenabhängige Parameter der Abstand der Korrespondenzen auf der Epipolargeraden. Dieser Abstand wird auch Disparität genannt. Der Suchbereich für die Disparität entlang der Geraden ist nur auf einer Seite durch den Epipol begrenzt. In der Realität kann jedoch

häufig eine maximale und minimale Tiefe der Szene abgeschätzt werden, wodurch der Suchbereich auf das Intervall d auf der Epipolargeraden eingeschränkt werden kann. Da für jeden Bildpunkt aus I_1 ein eigener Suchbereich in I_2untersucht werden muss, hat der gesamte Suchraum die Dimension $n \cdot m \cdot d = O\left(n^3\right)$.

9.5 Projektiver Dreibildfall: Trifokaltensor

Die Orientierung dreier Kameras kann im projektiven Raum durch den Trifokaltensor τ beschrieben werden. Es handelt sich dabei um eine $3 \times 3 \times 3$-Matrix, deren Notation mit $\tau = [T_1, T_2, T_3]$ definiert ist, wobei die drei T_i jeweils 3×3-Matrizen sind. Die wichtigste Eigenschaft des Trifokaltensors besteht in den sogenannten Trilinearitäten, die das Verhältnis zwischen korrespondierenden Punkten x, x' und x'' und Linien l, l' und l'', auf denen die Punkte liegen, beschreiben:

$$\begin{aligned}
&(1) \quad \left(l'^T \left[T_1, T_2, T_3\right] l''\right) [l]_\times = 0^T \\
&(2) \quad l'^T \left(\sum_i x_i T\right) l'' = 0 \\
&(3) \quad l'^T \left(\sum_i x_i T_i\right) [x'']_\times = 0^T \\
&(4) \quad [x']_\times \left(\sum_i x_i T_i\right) l'' = 0^T \\
&(5) \quad [x']_\times \left(\sum_i x_i T_i\right) [x'']_\times = 0_{3\times 3}
\end{aligned} \tag{41}$$

Besonders die Gleichung 41.5 erlaubt es, bei einem bekannten Korespondenzpaar die Position im dritten Bild oder analog zur Fundamentalmatrix die entsprechenden Epipolarlinien (Gleichungen 41.2-4) exakt zu berechnen. Aus diesem Tensor können drei projektiv verzerrte Projektionsmatrizen P_1, P_2 und P_3 abgeleitet werden, wenn angenommen wird, dass P_1 kanonisch ist. Details dazu sind in [22] beschrieben. Diese Projektionsmatrizen ermöglichen die Berechnung der Fundamentalmatrizen für jedes Kamerapaar:

$$F_{ij} = [P_j C_i]_\times P_j P_i^+ = [e_{ji}]_\times P_j P_i^+ \tag{42}$$

Hierbei ist P^+die pseudoinverse Projektionsmatrix von P und $[e_{ji}]_\times$ das schiefsymmetrische Produkt des Epipols von P_i in P_j. Allerdings sind diese Projektionsmatrizen projektiv verzerrt, weswegen eine originalgetreue Triangulation aus diesen Matrizen in der Regel nicht möglich ist. Eine projektive Triangulation und Rückprojektion allerdings ist sehr wohl möglich.

Eine weitere Eigenschaft der Trifokaltensors betrifft die Epipole der Fundamentalmatrizen, die aus dem Tensor berechnet werden. Der Tensor spannt zwischen den Epipolen analog zur Epipolarebene eine Trifokalebene auf (vgl. Abbildungen 6 und 8). Die durch diese Epipole bestimmten Fundamentalmatrizen erfüllen folgende Eigenschaft:

$$e_{23}^T F_{21} e_{13} = e_{31}^T F_{32} e_{21} = e_{32}^T F_{31} e_{12} = 0 \tag{43}$$

Diese Fundamentalmatrizen werden kompatibel genannt. Dabei ist zu beachten, dass paarweise berechnete Fundamentalmatrizen von drei Kameras nicht zwingend kompatibel sind, sondern nur, wenn sie über den Trifokaltensor abgeleitet werden, weil die Fundamentalmatrizen über projektive Mehrdeutigkeiten verfügen [22].

Der Trifokaltensor besteht aus 27 Elementen, besitzt aber nur 18 Freiheitsgrade. Er kann entweder direkt aus den Kameramatrizen oder projektiv aus sechs korrespondierenden Punkten in drei Bildern berechnet werden, wenn maximal drei dieser Punkte auf derselben Ebene im Raum liegen [62]. Befinden sich die Projektionszentren auf einer Linie, kann die Trifokalebene nicht bestimmt werden und eine Berechnung des Tensors aus Bildkorrespondenzen ist nicht möglich.

Stehen mehr Punktkorrespondenzen zur Verfügung, kann der Tensor robust parametrisiert werden. Dazu gibt es die Möglichkeit, entweder den algebraischen Fehler aus der Gleichung 41.5 oder den geometrischen Fehler aus der Rückprojektion der triangulierten

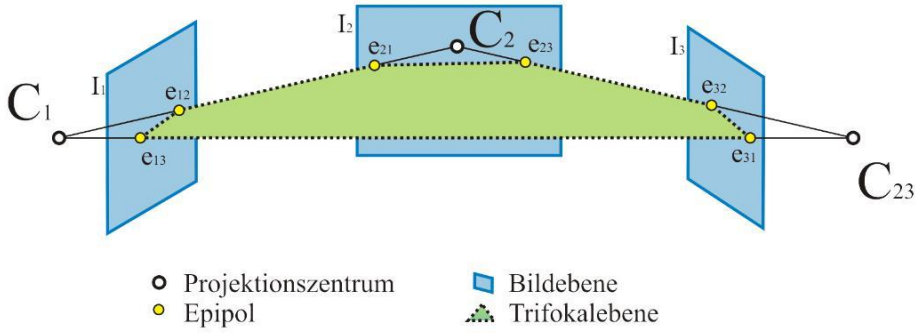


Abbildung 8: Trifokalebene

Raumpunkte zu minimieren. Das Rauschen in der Korrespondenzmessung wird nach einer Ausreißerfilterung als gaußverteilt angenommen, weswegen im Folgenden der Empfehlung aus [22] entsprochen wird, in diesem Fall den geometrischen Fehler zu minimieren.

9.6 Kalibrierter Zweibildfall: Elementarmatrix

Die Elementarmatrix E ist eng verwandt mit der Fundamentalmatrix. Die Epipolarbedingung (Gleichung 39.3) gilt auch bei der Elementarmatrix für alle korrespondierenden Punkte u und u' bei bekannten intrinsischen Kalibrierungen Kund K':

$$\begin{aligned} u &= K^{-1}x \\ u' &= K'^{-1}x' \\ u'^{T}Eu &= 0 \end{aligned} \tag{44}$$

Daher gilt auch bei der Elementarmatrix analog zur Fundamentalmatrix:

$$det\,(E) = 0 \tag{45}$$

Bei einer SVD-Dekomposition beträgt damit der kleinste Singulärwert Null. Zusätzlich gilt jedoch, dass die beiden anderen Singulärwerte von E gleich sind, was gleichbedeutend ist mit:

$$2EE^{T}E - spur\left(EE^{T}\right)E = 0 \tag{46}$$

Eine Elementarmatrix ist nur abhängig von der Rotation R und Translation t der Kamera P und kann bei bekannter Kamerakalibrierung direkt berechnet werden:

$$\begin{aligned} P &= K\,[R\,|t] \\ E &= t_{\times}R \end{aligned} \tag{47}$$

Umgekehrt können aus einer Elementarmatrix die Rotation und Translation bis auf einen Skalierungsfaktor wieder extrahiert werden. Eine genaue Beschreibung zur Extraktion dieser Parameter ist in [28, 22] beschrieben.

Es gibt viele Ansätze zur Berechnung der Elementarmatrix ausschließlich aus Bildkorrespondenzen. Eine Übersicht dieser Algorithmen kann in [54] nachgelesen werden. Der hier verwendete 5-Punktealgorithmus von [45, 61] ist interessant, da er besonders robust ist [54] und die benötigten fünf Korrespondenzen sogar auf einer Ebene liegen dürfen [50]. Allerdings hat dieser Algorithmus einen gravierenden Nachteil, da er auf die Bestimmung der Nullstellen eines Polynoms zehnten Grades basiert. Dieses Polynaom kann jedoch bis zu zehn reele Nullstellen haben und die richtige Lösung ist daher unter diesen bis zu zehn Kandidaten nicht eindeutig bestimmbar.

Teil III
Automatische 3D-Rekonstruktion

Die Generierung hochauflösender 3D-Modelle ist eine komplexe Aufgabe, die in mehrere Teilaufgaben untergliedert werden kann: Erstellung geradentreuer Bilder, Berechnung von Punktmerkmalen, synchrones Multikameratracking, Kamerapfadschätzung mit Bündelblockausgleich, Normalbilderstellung, dichte Korrespondenzsuche in Stereobildern und Triangulation der geschätzten Raumpunkte aus verschiedenen Zeitabschnitten zu einem einheitlichen Modell. Alle Module müssen zuverlässig und fehlertolerant arbeiten, um die Integration der Ergebnisse in den jeweils folgenden Teilbereich zu ermöglichen. Die Verknüpfung dieser einzelnen Teilaufgaben zu einer automatischen 3D-Modellgenerierung ist in Abbildung 9 skizziert.

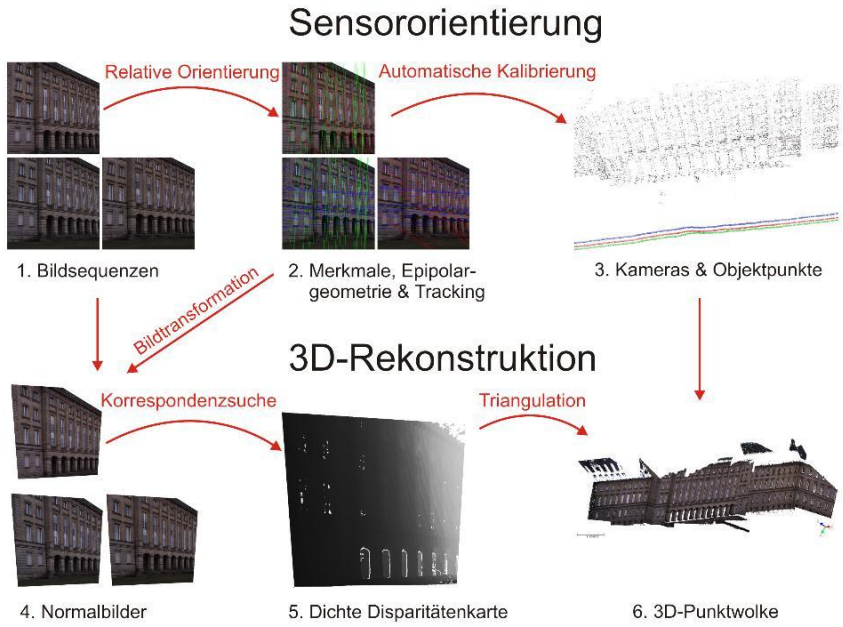


Abbildung 9: Prinzip der automatischen 3D-Modellgenerierung

Basis für eine hochgenaue Rekonstruktion ist die exakte Lokalisierung markanter Bildpunkte und ihre zuverlässige Zuordnung über mehrere Bilder. Dafür wird das Lokalisierungsverfahren von Förstner [14] um eine hochgenaue Subpixellokalsierung erweitert, die eine leichte Verbesserung der Arbeit von [53] beinhaltet. Diese Punkte werden mit der zuverlässigen Zuordnungstechnik der Arbeit von Lowe [39] verknüpft, um die gewünschte Robustheit und Genauigkeit zu erhalten.

Das Verfolgen markanter Bildpunkte in mehreren synchronen Videoströmen stellt besondere Anforderungen an die Trackingtechnik, die von klassischen KLT-Techniken nicht erfüllt werden. Zum einen müssen mehrere Datenströme gleichzeitig verarbeitet werden, zum anderen muss ein Bild nicht nur mit dem jeweils zeitlich folgenden verglichen werden, sondern auch mit den zwei Bildern der benachbarten Kameras auf dem Rahmen, was eine Optimierung per Sequenzialisierung sehr schwierig macht. Es wird basierend auf der Arbeit von [24] gezeigt, wie durch die geschickte Verwendung von Trifokaltensoren der Suchraum so eingeschränkt werden kann, dass die Merkmale, obwohl sie in mehr Bildern gefunden werden müssen, durchschnittlich länger verfolgt werden können als mit der KLT-Technik. Die Punktmerkmale können durch eine fixe Kamerageometrie nicht nur validiert, sondern sogar teilweise vorhergesagt werden, was eine starke Einschränkung der Position für mögliche Punktmerkmalskandidaten im Bild, die Überbrückung eines kurzen Verschwindens von Punkten sowie die Filterung vieler Fehldetektionen ermöglicht.

Die relative Orientierung der Aufnahmepositionen zweier aufeinander folgender Bilder

kann nicht immer eindeutig bestimmt werden (vgl. Abschnitt 9.6 und [61]). Da sich auf einem Rahmen montierte Kameras jedoch nicht unabhängig voneinander bewegen können, können diese Mehrdeutigkeiten aufgelöst und ungenaue Orientierungen gefiltert werden [54]. Gleichzeitig bietet ein fester Rahmen die Möglichkeit, den unbekannten und variablen Skalierungsfaktor zwischen zwei zeitlich verschiedenen Aufnahmepositionen global zu bestimmen. Die hochauflösenden und zeitlich dichten Bildsequenzen nehmen große Teile der untersuchten Objekte sehr oft auf. Diese hohe Redundanz bei der Objekterfassung soll verwendet werden, um die Robustheit und Genauigkeit der Kamerapfadschätzung per Bündelausgleichsrechnung zu steigern.

Um drei Bilder effektiv bearbeiten zu können, muss das Konzept der Normalbilder auf drei Bilder erweitert werden. Bilden die drei Kameras ein möglichst rechtwinkliges Dreieck kann eine Bildtransformation berechnet werden, die bezogen auf ein Referenzbild ein Bild horizontal mit Zeilenkongruenz und ein Bild vertikal mit Spaltenkongruenz erzeugt. Es wird ein lineares Rektifizierungsverfahren basierend auf den Arbeiten von [67, 25] vorgestellt und weitere Verbesserungen für die Wahl freier Parameter werden beschrieben.

Die dichte Korrespondenzsuche erzeugt im Zweibildfall ein grundsätzliches Problem: Liegen regelmäßige Strukturen parallel zur Verbindungslinie der Projektionszentren (vgl. Abbildung 6), kann auch eine Betrachtung der umliegenden Bildstrukturen wegen ihrer Regelmäßigkeit die Mehrdeutigkeiten nicht auflösen. Erschwerend kommt hinzu, dass regelmäßige Strukturen im vom Menschen gemachten Umfeld sehr häufig vorkommen und nicht ignoriert werden können. Wird ein drittes Bild zur Berechnung hinzugezogen, dessen Kameraprojektionszentrum nicht kollinear mit den anderen beiden ist, ändert sich diese Situation grundlegend: Zu jedem Korrespondenzkandidaten im zweiten Bild kann genau ein Punkt im dritten Bild berechnet werden, der dieser Korrespondenz entspricht. Dadurch können viele Mehrdeutigkeiten, die bei der Betrachtung eines Bildpaares auftreten würden, aufgelöst werden und die Korrespondenzsuche kann auch bei schwierigen regelmäßigen Strukturen wie Ziegelsteinen zuverlässige Ergebnisse liefern. Da die Berechnung dichter Korrespondenzkarten recht speicher- und zeitintensiv ist, werden die in [23] beschriebenen Erweiterungen und Verbesserungen des sehr universellen Korrespondenzsuchverfahren von [26, 27] weiterentwickelt und um eine akkurate Supixelapproximation erweitert, die auf der Arbeit von [15] aufbaut und diese bezüglich der Auswahl der Einflussregion verfeinert.

In einem abschließenden Schritt werden aus den berechneten Kamerapositionen und Bildkorrespondenzen mittels Triangulation 3D-Raumpunkte erzeugt. Dabei muss darauf geachtet werden, Duplikate desselben Objekts aus unterschiedlichen Bildern zu vermeiden oder zumindest zu erkennen, um das 3D-Modell mit möglichts wenig redundanten Informationen zu speichern.

Die algorithmische Analyse der Daten soll vollautomatisch erfolgen, es soll also möglichst kein Vorwissen über die aufgenommene Szene spezifiziert werden und die Algorithmen müssen sich automatisch auf geeignete Parameter einstellen oder mit Parametern arbeiten, die idealerweise für alle Anwendungen gültig sind. Einzige Ausnahmen bilden die innere Kalibrierung der verwendeten Kameras, die Verzeichnungsparameter und die Basislänge, die als gegeben vorausgesetzt werden.

Zur Aufnahme geeigneter Bildsequenzen wurde ein Gesamtsystem konzipiert, das aus vier Hardwareteilen besteht:

1. Drei FireWire800-Kameras liefern die Bilddaten. Die Auflösung der Kameras liegt bei maximal 2448x2048 Punkten. Die Bildwiederholrate liegt für diese Auflösung bei maximal 16 Vollbildern pro Sekunde. Die zeitliche Synchronisierung der Kameras erfolgt über Software, der zeitliche Versatz liegt unter 1ms. Da die Kameras über die FireWire800-Kabel auch ihre Stromversorgung erhalten, werden keine weiteren Kabel benötigt.

2. Die Kameras werden auf einem stabilen Aluminiumprofilrahmen befestigt (vgl. Abbildung 10), wo sie in einem rechtwinkligen, gleichseitigen Dreieck mit variabler Basislänge von 10 bis 90cm angeordnet sind. Für kleinere Objekte und Aufnahmen in engen Räumen wurde ein kleinerer Handrahmen mit Basislängen von 6 bis 25cm entworfen.

Abbildung 10: Kamerasystem mit Rahmen und Touchscreen und Aufnahmerechner

3. Die Aufnahme der Bilddaten erfolgt mit einem PC, der drei unabhängige FireWire800-Kanäle und einen Hardware-RAID-Controller besitzt, um die anfallenden Datenraten von max. 220 MB/s speichern zu können. Die Bedienung erfolgt über einen Touchscreen, der am Rahmen befestigt werden kann.

4. Um das System unabhängig vom Stromnetz betreiben zu können, wird ein 700W-Sinuskonverter mit einem 65Ah-Bleigelakku betrieben. Diese Komponenten sind für hohe konstante Stromflüsse optimiert, wodurch eine stromnetzunabhänige Aufnahmezeit von mehr als vier Stunden gewährleistet werden kann.

Der Prototyp kostete ca. 10000 Euro, wobei ein Anteil von 60% auf die Kameras und die Objektive entfiel.

10 Verfolgen markanter Punktmerkmale

Das System bestimmt die Kamerabewegung ausschließlich aus Bildkorrespondenzen markanter Merkmale. Durch das Tracking sollen diese Merkmale gefunden und verfolgt werden. Ferner sollen die Daten der drei Videokameras des Systems gleichzeitig verarbeitet werden und nicht nur Korrespondenzen von einem Bild zum nächsten (zeitliches Tracking), sondern auch zwischen den drei Kameras zum selben Zeitpunkt (räumliches Tracking) gefunden werden. Das Aufnahmesystem gewährleistet dabei, dass die Videos synchronisiert sind und die Bildtripel aus den drei Videos exakt zum selben Zeitpunkt aufgenommen wurden.

Standardtechniken aus Abschnitt 7.1 sind zwar prinzipiell in der Lage, das Tracking sowohl zeitlich als auch räumlich durchzuführen, die Algorithmen sind jedoch darauf optimiert, nur ein Video zu verarbeiten, bei dem sich ein Bild vom nächsten nur sehr geringfügig unterscheidet. Daher führt das räumliche Tracking mit seinen stärkeren perspektivischen Verzerrungen dazu, dass viele Merkmale nicht korrekt zugeordnet werden können. Es hat sich gezeigt, dass Qualität und Quantität dieser Merkmale nicht für eine zuverlässige Schätzung der Kamerabewegung ausreichen. Daher wird hier eine völlig neue Trackingtechnik vorgestellt, die folgende Kriterien erfüllt:

1. Genauigkeit: Die Bildposition der gefundenen Merkmale muss sich auf dieselbe Referenz im Raum beziehen. Da Ecken markante Bildpunkte sind und in der Regel Aliaseffekte im Pixelraster verursachen, sollte der Lokalisationsalgorithmus der Merkmale fähig sein, die Position des Merkmals im Subpixelraster zu approximieren.

2. Eindeutigkeit: Jedes Merkmal sollte so beschrieben werden, dass es auch bei leichten perspektivischen Verzerrungen in den Bildern immer eindeutig wiedergefunden werden kann.

3. Synchronität: Jedes Merkmal sollte sowohl in den drei Kameras eines Bildtripels als auch in drei aufeinander folgenden Bildtripeln verfolgt worden sein. Dadurch verringert sich zwar der Bildbereich, in dem Merkmale überhaupt gefunden werden können,

jedoch können auch erst dadurch hilfreiche Bedingungen für die Pfadextraktion eingeführt werden, die in Kapitel 11 benötigt werden.

4. Robustheit: Die Technik sollte Fehlzuordnungen und schlechte Positionierungen selbstständig erkennen und die entsprechenden Fehler eleminieren.

5. Projektive Validierung: Der Trifokaltensor erlaubt eine Beschreibung der perspektivischen Kamerarelationen über drei Positionen. Korrekt zugeordnete Merkmale in den Bildern dieser drei Kameras erfüllen die Trilinearitäten aus Gleichung 41 und können somit validiert werden.

6. Gerichtete Suche: Der Trifokaltensor erlaubt ferner, dass der Suchraum eines Merkmals in den anderen Bildern stark eingeschränkt wird. Wurde ein Merkmal schon in zwei Bildern gefunden, kann die gültige Position im dritten Bild sogar explizit angegeben werden.

7. Bewegungserkennung: Eine Voraussetzung der Pfadextraktion über die Zeit ist es, dass die Szene sich nicht bewegt. Leider ist diese Voraussetzung durch Wind und Verkehr häufig in Teilbereichen des Bildes nicht erfüllt. Diese beweglichen Bereiche müssen daher erkannt und herausgefiltert werden, um Fehler bei der Pfadextraktion zu vermeiden.

Hierbei ist zu beachten, dass die Kriterien 5, 6 und 7 ausreißerfreie Punktkorrespondenzen benötigen, um den Trifokaltensor zu bestimmen. Mit diesem Tensor können dann die übrigen Punkte validiert werden. Daher ist ein zweistufiges Verfahren nötig, um zunächst eine kleine Untermenge von sehr guten Korrespondenzen zu finden. Aus diesen guten Korrespondenzen kann in der zweiten Stufe ein Modell für die schwieriger zu findenden Korrespondenzen generiert werden. Leider erhöht sich der Rechenaufwand dadurch erheblich, weswegen Echtzeitanwendungen mit dieser Technik zur Zeit noch nicht möglich sind.

Die Verwendung des Systems zur Rekonstruktion von Bauwerken verstärkt ein prinzipielles Problem: Der Trifokaltensor kann nicht aus Punktkorrespondenzen bestimmt werden, die auf einer Ebene liegen. Dies ist jedoch im Bereich Architektur häufig der Fall, was zur Folge hat, dass erkannt werden muss, wann der Trifokaltensor degeneriert ist und eine Homographieberechnung, die ausschließlich das Verhalten von Punkten auf einer Ebene beschreibt, vorgezogen werden sollte.

Dieses Kapitel unterteilt sich in Abschnitte zur Bildkorrespondenzlokalisation und -suche sowie zur Filterung der korrespondierenden Merkmale mittels Trifokaltensoren. Abschließend werden Ergebnisse für das Tracking vorgestellt und mit dem Standard-KLT verglichen.

10.1 Bildkorrespondenzen

Die Grundtechniken für jedes Trackingverfahren bestehen aus der Lokalisierung der markanten Punkte, der Beschreibung dieser Punkte und dem Finden von Korrespondenzen. Dazu werden die theoretischen Grundlagen aus Abschnitt 6 angewandt und erweitert, indem ein Verfahren zur Verwendung der Förstnerpunkte mit SIFT-Deskriptoren beschrieben wird. Das ursprüngliche Zuordnungsverfahren von SIFT-Deskriptoren [39] wird um eine symmetrische Komponente erweitert. Anschließend wird die Subpixelgenauigkeit der Förstnerlokalisation im Vergleich zur SIFT-Lokalisation untersucht.

10.1.1 Koppelung von Förstnerpunkten mit SIFT- Deskriptoren

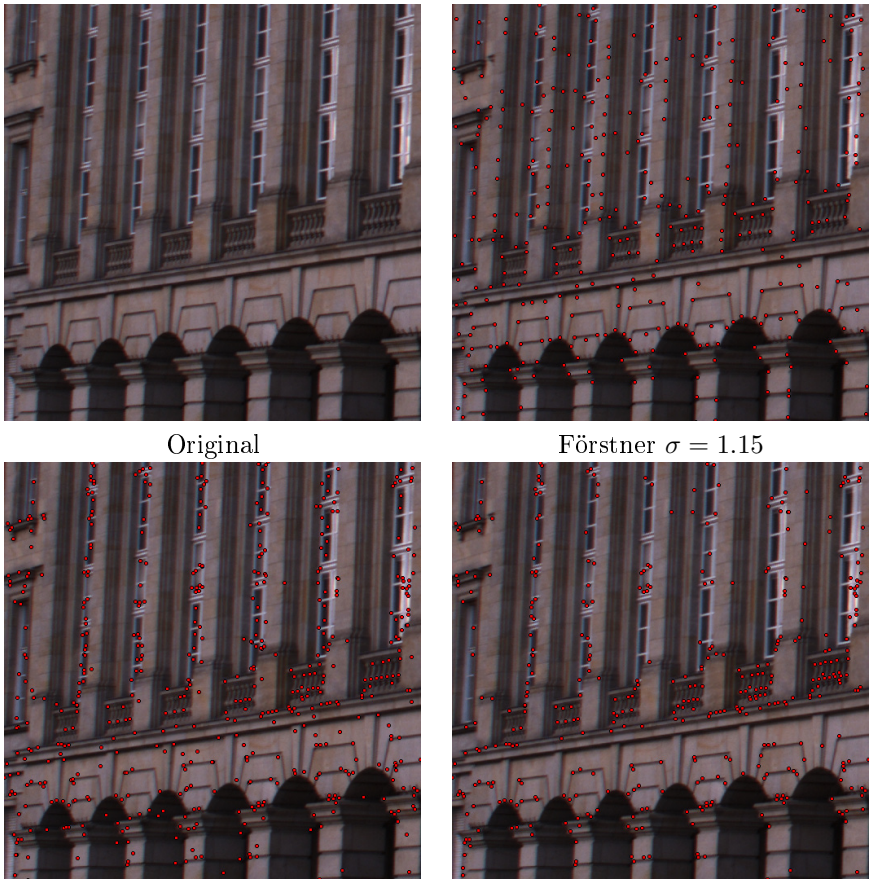
Die SIFT-Technik ist invariant gegenüber Skalierung. Dies wird erreicht, indem das Bild sukzessive über eine Skalenpyramide verkleinert wird und Punktmerkmale über diese Skalen separat extrahiert werden. Leider hat dies aber auch zur Folge, dass Merkmale in sehr großen Skalen nicht so genau lokalisiert werden können wie in niedrigeren Skalen.

Ein Beispiel dafür ist in Abbildung 11 gegeben. Die obere Zeile zeigt rechts den originalen Bildausschnitt und links die extrahierten Förstnerpunkte mit $\sigma = 1.15$. In der unteren Zeile sind auf der linken Seite ein Bildausschnitt und die extrahierten SIFT-Punkte mit einem

Oktavensigma von 1.15 gezeigt, rechts daneben wurden mit denselben Einstellungen die SIFT-Punkte nur für die erste Skala berechnet. Bei den SIFT-Punkten liegt ein großes Problem in der Mehrfacherkennung desselben Merkmals in verschiedenen Skalen. An den Ecken der Kapitele und an den Fugen in den Säulenbögen ist auf dem unteren linken Bild zu erkennen, dass es zu einer Gruppenbildung der Merkmale in den Ecken kommt. Die Punkte liegen so nah beieinander, dass eine eindeutige Lokalisierung des eigentlichen Merkmals sehr schwer fällt. Dieser Effekt ist auf dem unteren rechten Bild, bei dem nur die erste Skala der SIFT-Features verwendet wurde, in geringerem Maße und bei den Förstnerpunkten gar nicht vorhanden.

Bei den SIFT-Varianten kommt es zu vielen Punktdetektionen in Bereichen ohne erkennbare Struktur, beispielsweise auf den Fensterflächen und entlang der Fugen der Klinkersteine. Dies ist darauf zurückzuführen, dass SIFT sogenannte Blobs (Strukturänderung in einem punktförmigen Bereich) sucht und damit die Mitte der hohen Fenster findet, wenn die Fensterhöhe mit dem Radius des Suchbereiches korreliert. Diese Detektion der Flächenmitte ist jedoch nicht zuverlässig, da durch die perspektivische Verzerrung die Mitte des abgebildeten Fensters nicht seiner tatsächlichen Mitte entspricht und stark vom Betrachtungswinkel abhängt. Bei den Förstnerpunkten mit $\sigma = 1.15$ wurden die Ecken sehr zuverlässig lokalisiert und zeigen an den Säulenbögen ein gleichmäßigeres Verhalten als SIFT.

Daher werden zur Punktextraktion Förstnerpunkte mit kleinem σ bevorzugt. Bei diesem Ansatz wird auf die Skaleninvarianz verzichtet, allerdings ist sie bei der Videoanalyse nicht von großer Relevanz, da durch die dichte Bildfolge von einem Bild zum nächsten kaum Größenunterschiede der Szene zu erwarten sind.



Original | Förstner $\sigma = 1.15$

SIFT $\sigma = 1.15$ | SIFT mit fixer Skala

Abbildung 11: Vergleich von Punktmerkmalen

Die SIFT-Deskriptoren sind prinzipiell unabhängig von dem Blob-Detektor, benötigt wird lediglich der Einflussbereich des Deskriptors um den Bildpunkt, welcher bei den SIFT-

Punkten implizit über die Skala gegeben ist. Bei Förstnerpunkten muss diese Skala aus dem σ berechnet werden, was vom Prinzip her genau der Skalaberechnung der SIFT-Punkte entspricht. Allerdings wird beim SIFT-Deskriptor ein konstanter Vergrößerungsfaktor verwendet. Daher entspricht der Radius der Einflussregion des Deskriptors in Bildpunkten zwölf mal σ .

10.1.2 Zuordnen von Deskriptoren

Die Ähnlichkeit zweier jeweils 128 Elemente großer Deskriptoren a und b kann über den euklidischen Abstand berechnet werden:

$$cost_{SIFT}(a,b) = \sqrt{\sum_{i=1}^{128} (a_i - b_i)^2} \tag{48}$$

Allerdings kann durch das unbekannte Rauschverhalten im Bild nur selten ein globaler Schwellwert angegeben werden, der zuverlässig alle gültigen Zuordnungen findet. Die in [39] vorgeschlagene Technik zur Bestimmung einer gültigen Zuordnung berechnet sich aus dem Verhältnis der besten zu den zweitbesten Kosten:

$$\frac{cost_{bester}}{cost_{zweitbester}} < thres_{SIFT} \tag{49}$$

Diese Berechnung ist unabhängig vom Bildrauschen, geeignete Werte für den Schwellwert sind $0.5 \leq thres_{SIFT} \leq 0.8$. Allerdings kann es sein, dass der gewählte Punkt im zweiten Bild einen anderen Punkt im ersten als sein Minimum betrachtet, und es kommt zu Doppelzuordnungen. Diese können zwar nachträglich gefiltert und verworfen werden, jedoch stellt sich die Frage, ob beide Zuordnungen oder nur eine als ungültig deklariert werden. Daher wird dieses Zuordnungsverfahren um eine symmetrische Komponente erweitert: Zunächst werden die Kosten für alle Deskriptorkombinationen aus dem ersten und zweiten Bild berechnet und in einer $n \times m$-Kostenmatrix aufgetragen, wobei die Zeilen dieser Matrix die Deskriptoren des ersten Bildes und die Spalten die Deskriptoren des zweiten Bildes indizieren. Kandidaten für gültige Zuordnungen sind Punktpaarungen, bei denen das Minimum einer Zeile gleichzeitig das Minimum der entsprechenden Spalte ist. Diese Kandidaten müssen zusätzlich die Bedingung der Gleichung 49 für die jeweils zweitbesten Werte in der Zeile bzw. Spalte erfüllen, die natürlich an unterschiedlichen Positionen liegen.

10.1.3 Subpixellokalisation

Die Subpixellokalisation der Förstnerpunkte erfolgt wie in Abschnitt 6.1.1 erwähnt über eine Paraboloidapproximation der Operatorstärke w_i (Gleichung 10) an der Merkmalsposition x_0 und der acht Werte der Rasternachbarn. Da eine Non-Maxima-Filterung stattfand, ist sichergestellt, dass die Position x_0 ein lokales Maximum ist und der Paraboloid ein lokales Maximum in der Nähe dieser Position haben muss. Die Werte werden vor der Paraboloidbestimmug auf den Maximalwert w_0 normiert. Das Paraboloid ist über sechs Parameter definiert:

$$a\,(x_i^{x\,\mathbf{2}}) + b\,(x_i^{y\,\mathbf{2}}) + cx_i^x x_i^y + dx_i^x + ex_i^y + f = w_i \tag{50}$$

Ein elliptisches Paraboloid liegt vor, wenn die Vorzeichen von a und b gleich sind. Die Position des lokalen Maximums dieses Paraboloiden definiert die Verschiebung im Subpixelbereich und berechnet sich aus:

$$\begin{aligned} sub^x &= \tfrac{2bd-ce}{c^{\mathbf{2}}-4ab} \\ sub^y &= \tfrac{2ae-cd}{c^{\mathbf{2}}-4ab} \end{aligned} \tag{51}$$

Allerdings ist der Paraboloid durch neun Werte überbestimmt und es kann passieren, dass eine Kleinste-Quadrate-Lösung zu einem hyperbolischen Paraboloid (a und b haben

Algorithmus 1 Subpixelapproximation von Förstnerpunkten

1. Bestimme die Punktstärken in einem 3×3-Gebiet um das Merkmal an Position x_0
2. Normiere die Werte auf die Stäke an Position x_0
3. Bestimme die Parameter nach Gleichung 50 des Paraboloiden aus diesen neun Werten mittels SVD
4. Bestimme die Extremstelle nach Gleichung 51
5. Ist die maximale Iterationsstufe noch nicht erreicht?
   (a) Ja: Ist ein Versatz in mindestens einer Richtung größer als ein halber Punkt?
      i. Ja: Verschiebe x_0 um einen Punkt in diese Richtungen (Relokalisation) und gehe zu Schritt 1
      ii. Nein: Gib die Subpixelposition inkl. Relokalisierung zurück
   (b) Nein: Relokalisierung ist nicht stabil: Verwirf den Punkt

ungleiche Vorzeichen) mit einem Sattelpunkt in der Nähe von x_0 führt. In diesem Fall ist eine Paraboloidapproximation nicht sinnvoll und der Punkt wird verworfen.

Ein zweiter Aspekt sind stark verschobene Paraboloiden. Bei neun Werten kann ein Ungleichgewicht auftreten, und das Maximum des Paraboloiden um mehr als einen halben Bildpunkt von der Position x_0 abweichen. Es hat sich gezeigt, dass besonders gut zu erkennende Merkmale diesen Effekt aufweisen, diese Punkte sollten daher nicht verworfen werden. Inspiriert durch die Subpixelapproximation des SIFT wird geprüft, ob sich die Subpixelposition in vier Iterationen stabilisiert, sich also unter einem halben Bildpunkt Versatz befindet. Die gesamte Subbixelaproximation ist in Algorithmus 1 zusammengefasst.

10.1.4 Ergebnisse

Um eine Abschätzung der Subpixelpositionierungsgenauigkeit von Förstnerpunkten durchführen zu können, wird ein teilsynthetisches Testszenario verwendet: Eine Szene wird zweimal vom selben Standpunkt aus fotografiert, um ein realistisches Sensorrauschen zu modellieren. Von diesem Bild wird nur eine kreisförmige 512×512-Region verwendet, die mit einem 100 Punkte breiten, schwarzen Rand umgeben wird, um Randeffekte zu vermeiden. Das zweite Bild wird ebenso beschnitten und dann durch eine Transformation im Subpixelbereich verschoben und neu interpoliert: Danach werden von diesen Bildern mittels der beschriebenen Techniken Punktkorrespondenzen gesucht. Die Positionsänderung der korrespondierenden Punkte wird nun mit der angewendeten Transformation verglichen. Die durchschnittliche Positionsabweichung vom Sollwert wird als Fehlermaß angegeben und Punktkorrespondenzen mit einem Abstand größer als 0.5 Bildpunkte in x- oder y-Richtung als Ausreißer klassifiziert. Zusätzlich wird die Standardabweichung dieser Positionierungsfehler berechnet. Zu Vergleichszwecken wurden diese Tests mit der originalen Implementierung von Lowe (Weblink in Anhang B) durchgeführt.

Die Transformationen umfassen vier Typen: Translation, Rotation, perspektivische Verzerrung und Skalierung. Die Translationstests erfolgen in 0.1-Punkt-Schritten im Intervall $[-1; 1]$ entlang der x- und y-Achse. Die Rotationstests wurden in 5°-Schritten durchgeführt. Für die perspektivische Verzerrung wurden die Ecken des Bildes in 2-Punkt-Schritten im Intervall $[-20, 20]$ trapezförmig verschoben. Für die Skalierung wurden die Eckpunkte ebenfalls in 2-Punktschritten im Intervall $[-20, 20]$ nach außen bzw. nach innen verschoben.

In Abbildung 12 sind die Ergebnisse der Translationstests illustriert. Die Verschiebung ist in der $x - /y$-Ebene aufgetragen, der Positionierungsfehler bestimmt den Wert der z-Achse. Die linke Seite zeigt oben den Verschiebungsfehler der Förstnerpunkte und unten die dazugehörige Standardabweichung. Auf der rechten Seite ist oben die Fehlerdifferenz der SIFT-Punkte zu den Förstnerpunkten aufgetragen und unten die Differenzen der Standard-

abweichungen. Positive Zahlen zeigen, wieviel genauer die Lokalisierung der Förstnerpunkte im Vergleich zur SIFT-Lokalisierung war.

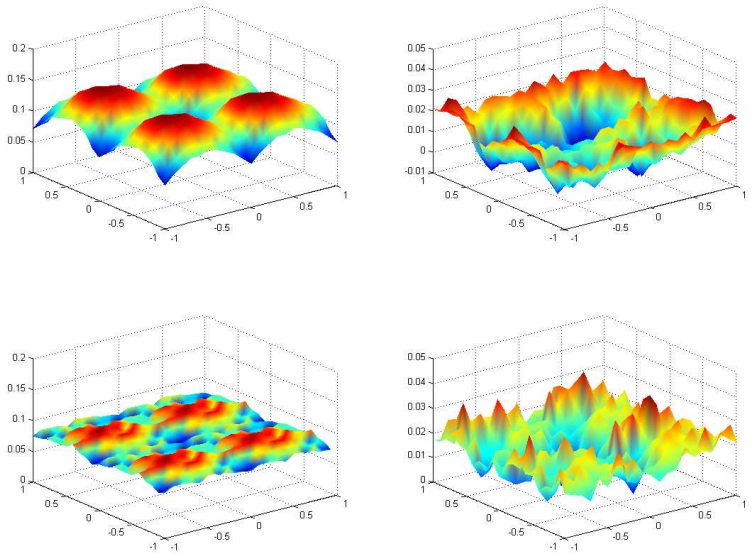


Abbildung 12: Positionierungsfehler im Subpixelbereich

	Abweichung	Tr	Ro	P	S
Förstner	Position max	0.1444	0.1370	0.1857	0.1805
	Position Ø	0.1141	0.1207	0.1311	0.1204
	σ max	0.0933	0.1031	0.1100	0.1148
	σ Ø	0.0816	0.0890	0.0905	0.0847
SIFT	Position max	0.1493	0.1462	0.1501	0.1419
	Position Ø	0.1249	0.1333	0.1334	0.1295
	σ max	0.1146	0.1162	0.1152	0.1109
	σ Ø	0.0997	0.1056	0.1057	0.1030
Verbesserung	Position Ø	0.0107	0.0125	0.0024	0.0091
Förstner zu SIFT	Position min	-0.0066	-0.0037	-0.0445	-0.0396
	Position max	0.0270	0.0247	0.0425	0.0465
	σ Ø	0.0181	0.0166	0.0142	0.0183
	σ min	0.0012	-0.0024	-0.0112	-0.0068
	σ max	0.0347	0.0294	0.0350	0.0363

Tabelle 2: Ergebnisse Positionierungsfehler

Der maximale Fehler bei der Translation tritt wie zu erwarten bei der maximalen Verschiebung von $\sqrt{\frac{1}{2}} \approx 0.7071$ auf und beträgt 0.1444 Bildpunkte (vgl. Abbildung 12 oben links und Tabelle 2). Dieser Restfehler hat eine Standardabweichung von 0.0933. Damit kann im Subpixelbereich im schlechtesten Fall eine Positionierungsgenauigeit des Förstneroperators von 0.2377 angenommen werden. Auffällig ist, dass bei kleinen Verschiebungen inbesondere die Positionierungsgenauigkeit der Förstnerpunkte besser ist als die der SIFT-Punkte, wie an den roten Rändern von Abbildung 12 oben rechts zu erkennen ist. Nur bei sehr starker Subpixelverschiebung schneiden die SIFT-Punkte um durchschnittlich 0.0066 Punkte besser ab, was in der Praxis jedoch nicht relevant ist. Generell steigt die Standardabweichung

der Förstnerpunkte mit der Zunahme der Positionierungsfehler nur marginal und liegt im Durchschnitt immer unter der der SIFT-Punkte. Dies zeigt, dass die Lokalisierung der Förstnerpunkte sehr viel stabiler und zuverlässiger ist.

Alle Ergebnisse für Translation (Tr), Rotation (Ro), perspektivische Verzerrung (P) und Skalierung (S) gehen aus Tabelle 2 hervor. Im unteren Teil sind die Verbesserungen der Förstnerpunkte gegenüber den SIFT-Punkten aufgeführt, wobei negative Zahlen eine Verschlechterung anzeigen. Insgesamt ist der Förstneroperator genauer und robuster als der SIFT-Operator, nur bei der perspektivischen Verzerrung entspricht sich die Genauigkeit der beiden Verfahren. Selbst bei kleinen Skalierungsänderungen ist der Förstneroperator genauer, da diese sehr kleinen Größenunterschiede durch den SIFT-Operator nicht korrekt detektiert werden können. Auch wenn die SIFT-Punkte in Bezug auf die Standardabweichung bei perspektivischer Verzerrung und Skalierung punktuell geringfügig stabiler sind, zeigt die durchschnittliche Verbesserung der Standardabweichung, dass die Positionierungsgenauigkeit des Förstneroperators besser ist als die des SIFT-Operators oder im Falle der perspektivischen Verzerrung zumindest gleich gut.

10.2 Trifokalfilter

Im Rahmen der Zuordnungsstrategie aus Abschnitt 10.1.2 wird eine n-zu-m-Zuordnung der Deskriptoren auf nicht sortierten Daten durchgeführt. Da n und m die Anzahl der markanten Bildpunkte in den jeweiligen Bildern sind, kann ohne Einschränkung der Allgemeinheit gesagt werden: $m = n$. Der Aufwand ist daher:

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} = O\left(n^2\right) \tag{52}$$

Auch das ursprüngliche Zuordnungsverfahren weist diesen quadratischen Aufwand auf, weil zu jedem der n Deskriptoren aus der ersten Menge alle m Deskriptoren aus der zweiten Menge untersucht werden müssen, um das Minimum zu finden. Dies ist für hochauflösende Bilder mit ca. 5000 markanten Punkten nicht praktikabel, da für die Zuordnungen der Deskriptoren mehr als 12.5 Millionen Prüfungsberechnungen durchgeführt werden müssen. Der Aufwand ließe sich durch Sortierung auf $O\left(n \log n\right)$ reduzieren, aber eine sinnvolle Vorsortierung der SIFT-Deskriptoren ist wegen der hohen Dimensionalität nicht durchführbar, da kein eindeutiges Sortierkriterium für 128-dimensionale Vektoren existiert (vgl. "*Curse of dimensionality*", [3]). Besteht allerdings die Möglichkeit, Bedingungen an den Aufenthaltsort des Merkmals zu formulieren, kann eine Filterung der Merkmale anhand der vorsortierten Bildpositionen erfolgen, was den Rechenaufwand drastisch reduziert.

Des weiteren kann nicht verhindert werden, dass ohne Bedingungen der Bildposition einige Fehlzuordnungen erfolgen. Insbesondere sich wiederholende Muster wie Fenster lassen sich ohne Einschränkung der Position nicht eindeutig zuordnen, da die Deskriptoren an Fensterecken desselben Fenstertyps idealerweise sogar gleich sein sollen.

Eine zusätzliche Bedingung ans System ist, dass die Punkte sowohl zwischen den einzelnen Kameras als auch von Bild zu Bild verfolgt werden müssen. Dadurch muss dasselbe Merkmal in mindestens sechs Bildern wiedergefunden werden, was zunächst sowohl die Robustheit als auch die Anzahl der Merkmale stark beeinträchtigt. Im vorgestellten System wird durch den festen Rahmen eine fixe Beziehung zwischen den drei räumlichen Kameras erzeugt. Diese Beziehung kann durch den Trifokaltensor ermittelt werden, der bis auf gewisse Toleranzen als unveränderlich angenommen wird. Dadurch können bei der Korrespondenzsuche geometrische Bedingungen an die Merkmale eines Bildtripels gestellt und so die Robustheit und Anzahl der korrespondierenden Merkmale erhöht werden, so dass die Zuordnung der Merkmale qualitativ und quantitativ bessere Ergebnisse liefert als eine unbedingte Zuordnung.

Aus drei aufeinander folgenden Bildern kann ebenfalls ein Trifokaltensor bestimmt werden, wenn die Kameras sich bewegt haben und vorab sechs Korrespondenzen bekannt sind, die nicht auf einer Ebene liegen. In diesem Fall können die Bedingungen nur nachträglich eingeführt werden, weil sie dynamisch aus einigen gültigen Korrespondenzen ermittelt werden

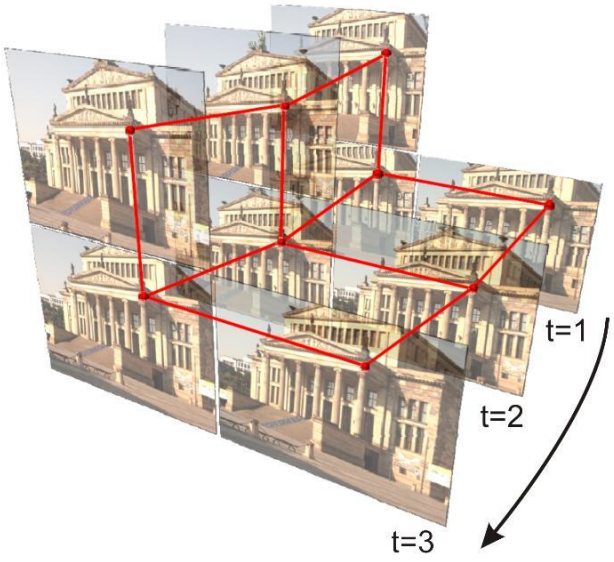


Abbildung 13: Netzwerk des Trifokalfilters

müssen. Danach ist es möglich, anhand eines aus vorläufigen Korrespondenzen bestimmten Trifokaltensors Bedingungen zu formulieren, die in einem zweiten Durchlauf die vorläufigen Zuodnungen validieren und zusätzliche Merkmale finden können, die beim ersten Durchlauf wegen Mehrdeutigkeiten verworfen wurden.

Werden für die zeitliche Verfolgung von Punkten nur solche verwendet, die vorher in den drei Kameras gefunden wurden, entsteht ein Netz aus neun Punkten, die alle dasselbe Merkmal im Raum beschreiben (vgl. Abbildung 13). Die Bedingungen an dieses Netzwerk werden Trifokalfilter genannt. Der Trifokalfilter ermöglicht eine Einschränkung der Korrespondenzsuche im Bild auf bestimmte Ausschnitte, im Folgenden Guided Matching genannt, und kann sogar wegen der Redundanz mancher Informationen Defekte im Netzwerk der Korrespondenzen reparieren, indem er fehlende oder verlorene Punktmerkmale künstlich neu erzeugt.

10.2.1 Guided Matching

Für das Guided Matching sind drei Fälle zu beachten. Zunächst kann der Suchraum im zweiten Bild mittels Fundamentalmatrix für jedes Punktmerkmal im Referenzbild auf einen dünnen Streifen beschränkt werden. Steht keine Fundamentalmatrix, sondern eine Homographie zur Verfügung, kann versucht werden, für jedes Merkmal im Primärbild eine kreisförmige Region im Sekundärbild zu finden, in der sich das Merkmal befinden muss. Im Dreibildfall werden durch die Eigenschaften des Trifokaltensors weitere Bedingungen eingeführt. Diese drei Fälle werden nun im Einzelnen erläutert.

Guided Matching mittels Fundamentalmatrix Ist die Fundamentalmatrix F für das Bilderpaar bekannt, kann als örtliches Filterkriterium die Epipolargeometrie verwendet werden. Für ein Merkmal am Ort x kann ein Toleranzbereich um die Epipolargerade $l = Fx$ definiert werden, innerhalb dessen sich die geometrisch gültigen Merkmale im zweiten Bild befinden dürfen (vgl. Abbildung 6). Der Toleranzbereich wird durch den maximalen Sampsonabstand (Gleichung 56) der Punkte definiert, die zur Berechnung der Fundamentalmatrix dienten. Dadurch ist sichergestellt, dass kein Punktmerkmal geometrisch schlechter als der schlechteste Punkt ist, der für die projektive Kalibrierung verwendet wurde. Diese geometrische Bedingung ist durch keine Kamerabewegung limitiert und kann auch gegenläufige Merkmale verfolgen, z. B. bei Drehungen um einen Punkt in der Szenenmitte, was bei der Filterung durch optischen Fluss [29] nicht uneingeschränkt möglich ist. Allerdings können Fehlzuordnungen entlang der Epipolargeraden nicht verhindert werden.

Guided Matching mittels Homographie Wenn die Kamera bei der Aufnahme stehen bleibt, nur rotiert oder nur eine planare Fläche aufnimmt, ist die Berechnung einer Fundamentalmatrix nicht möglich und eine Einschränkung der Positionen wie im vorherigen Abschnitt ausgeschlossen. In diesem Fall beschreibt eine Homographie H den Zusammenhang der zwei Bilder bzw. den Zusammenhang der Punkte auf der Ebene. Mit Hilfe dieser Transformation kann die Position x im ersten Bild direkt auf eine Position $x' = Hx$ im zweiten Bild transferiert werden, wobei die Toleranz für die Merkmale ein Kreis mit Radius r um den transferierten Punkt x' ist. Der Radius sollte analog zur Fundamentalmatrix aus dem maximalen Abstand der gültigen Zuordnungen abgeleitet werden, um wieder sicherzustellen, dass kein Punkt schlechter zur Homographie passt als der bisher am schlechtesten passende.

Guided Matching mittels Trifokaltensor Die Verwendung des Trifokaltensors, der den geometrischen Zusammenhang dreier Bilder beschreibt (vgl. Abschnitt 9.5), ermöglicht eine weitere Technik zur Bereichseinschränkung. Zum einen können die drei Fundamentalmatrizen für die drei möglichen Kamerapaare aus dem Tensor extrahiert werden, zum anderen kann mittels des Tensors aus einem korrespondierenden Paar die entsprechende Position des Merkmals im dritten Bild direkt bestimmt werden. Wie bei der Homographie wird innerhalb eines Radiuses um diese Position nach dem korrespondierenden Merkmal gesucht. Die Bedingungen an dieses Tripel sind über Gleichung 41 definiert, wodurch Fehlzuordnungen entlang einer Epipolarlinie ausgeschlossen werden können. Fehlerhafte Zuordnungen können nur noch dann auftreten, wenn die drei Punkte Gleichung 41 erfüllen und sich damit Merkmale an drei geometrisch gültigen Positionen befinden, die jedoch nicht zum selben Objekt gehören. Um solche Fehlzuordnungen zu vermeiden, dürfen die Merkmale eine gewisse Mindestähnlichkeit nicht unterschreiten. Ein Test nach Gleichung 49 ist nicht möglich, wenn es nicht mehrere Kandidaten im Zielbereich gibt. Um zufällige Fehlzuordnungen, die zwar geometrisch richtig sind, aber immens hohe Kosten mit sich bringen, zu vermeiden, werden die Kosten des Tripels gegen einen globalen Schwellwert von 255 getestet, was einer erlaubten durchschnittlichen Abweichung von zwei Zahlenwerten über den gesamten Vektor entspricht. Die so bestimmten Korrespondenzen sind praktisch frei von Ausreißern. Bei ausgiebigen Testläufen wurden nur dann vereinzelt Fehler gefunden, wenn die Kamera sich nicht bewegte und der Trifokaltensor degeneriert war.

Vermeidung von degenerierten Fällen Da auch der Trifokaltensor nicht aus Punkten auf einer Ebene im Raum berechnet werden kann, muss eine Strategie gefunden werden, die degenerierten Fälle zu erkennen und wieder auf eine Homographie zurückzugreifen. Hierfür wird die Kostenfunktion des robusten GASAC-Schätzers [55] - einer Weiterentwicklung von RANSAC [13]- um einen Homographiecheck erweitert: Mittels SVD wird aus den sechs Punkten, die zur Trifokaltensorberechnung verwendet wurden, eine überbestimmte Homographie berechnet und die Residuen dieser Punkte bestimmt. Lagen die Punkte nicht auf einer Ebene, entsteht mindestens ein Residuenfehler mit mehr als einem Bildpunkt. Sind jedoch alle Fehlerresiduen sehr klein, lagen die Punkte auf einer Ebene und der Trifokaltensor ist aus diesen Punkten nicht bestimmbar. In diesem Fall werden die Kosten des Trifokalfehlers um einen konstanten Term erhöht, der einem Abstand von fünf Bildpunkten entspricht, was dazu führt, dass Trifokaltensoren bevorzugt werden, die nicht aus einer degenerierten Konfiguration berechnet wurden. Gibt es ausschließlich degenerierte Konfigurationen, ist eine Homographieberechnung für die drei Bilder vorzuziehen. Die Kostenberechnung des Trifokaltensors für die GASAC-Bewertung ergibt sich aus Algorithmus 2. Die konstante Strafe im fünften Schritt wurde auf fünf Bildpunkte gesetzt. Somit ist es möglich die degenerierten Fälle zu vermeiden oder, wenn es nur degenerierte Fälle gibt, wenigstens zu erkennen.

Algorithmus 2 Trifokalkosten mit Homographieprüfung

1. Bestimme die Basiskosten aus dem Rückprojektionsfehler:
2. Trianguliere die Raumpunkte X^j aus den Korrespondenzen $\left\{x_1^j, x_2^j, x_3^j\right\}$ der Bilder 1,2 und 3 und aus den aus dem Trifokaltensor extrahierten Projektionsmatrizen P_1, P_2 und P_3.
3. Berechne die Rückposition $\bar{x}_i^j$
   (a) Berechne die Rückprojektionsfehler $err_{proj}\left(x_i^j, \bar{x}_i^j\right)$ für alle Punkte und berechne den Median err_{median} aus diesen Fehlern.
   (b) Bestimme einen maximal erlaubten Fehler für n Punkte mit 18 linearen Bedingungen nach [8] aus dem Median:
   $err_{max} = 2.79 \cdot 1.4826 \cdot (1.0 + \frac{5.0}{n-18}) \cdot err_{median}$
   (c) Begrenze die Rückprojektionsfehler auf err_{max} , indem alle Punkte mit einem höherem Fehler entfernt werden, und bilde den durchschnittlichen Rückprojektionsfehler für alle guten Korrespondenzen.
4. Für jedes Bildpaar des Tripels:
   (a) Berechne eine Homographie für die sechs Punktpaare der Bilder.
   (b) Berechne die Residuen. Bei kleinen Residuen bilden die sechs Korrespondenzen eine Homographie und keinen Trifokaltensor.
5. Sind alle Residuen kleiner Eins, addiere eine konstante Strafe auf die Basiskosten.

Guided Matching mittels Trifokalfilter Mit dem Trifokaltensor kann jetzt der Trifokalfilter für drei aufeinander folgende Bilder einer Videosequenz aufgebaut werden (Algorithmus 3). Dabei ist zu beachten, dass sich die Bildtripel bei der Videobearbeitung immer nur um ein Bild verschieben. Daher kann die Fundamentalfilterung aus Schritt 3c ab dem zweiten Bild entfallen, weil aus der vorherigen Iteration bereits gültige Punktkorrespondenzen für das erste Bildpaar bekannt sind.

Um diesen Filter auf drei synchrone Videos anwenden zu können, wird die Technik um einige Punkte erweitert: Zunächst werden Korrespondenzen für sämtliche Punktmerkmale eines zur selben Zeit aufgenommenen Bildtripels mit einem statischen Trifokaltensor für den Rahmen und Algorithmus 3 berechnet und gefiltert. Dieses Verfahren wird für drei zeitlich aufeinander folgenden Bildtripel wiederholt. Die so bestimmten Punkte erfüllen die Bedingungen an den Trifokaltensor des Rahmens und nur diese kommen daher für die Punktverfolgung in Betracht. Nun werden für jedes zeitliche Tripel die bereits räumlich korrespondierenden Punkte mit einem separaten einfachen Trifokalfilter nach Algorithmus 3 untersucht. Widersprechen sich Zuordnungen aus den unterschiedlichen Videos, sind also die räumlichen Tripelzuordnungen nicht konsistent, wird der Kandidat verworfen. Die so berechneten Punktkorrespondenzen erfüllen das Netzwerk wie in Abbildung 13 gezeigt.

10.2.2 Reparatur von Defekten

Stehen zwei korrespondierende Punkte und ein Trifokaltensor zur Verfügung, kann die Position im dritten Bild bestimmt werden, da die Positionsinformation redundant ist. Durch diese Redundanz können Lücken im Trifokalfilter aufgefüllt werden, was die Stabilität der verfolgten Punkte steigert, wenn diese durch Verdeckung kurzzeitig nicht zu sehen sind. Die so eingefügten Punkte werden virtuelle Merkmale genannt. Um spätere Fehlzuordnungen zu vermeiden, darf sich an dieser reparierten Stelle kein reales Punktmerkmal befinden.

Theoretisch ist es möglich, aus vier Punkten in zwei Bildern alle neun Punkte zu rekonstruieren. Allerdings wären diese Punkte sehr spekulativ und ihre Lokalisation durch

den Trifokaltensor zu ungenau. Daher wird empfohlen, maximal einen Defekt pro Tensor zu erlauben. Da jede Position die Bedingungen zweier Tensoren erfüllen muss, werden nur die virtuellen Positionen akzeptiert, die durch beide Trifokaltensoren bestätigt werden, da so gewährleistet ist, dass zumindest die Genauigkeit des räumlichen Trifokaltensors auch bei geringen Bewegungen des Rahmens eingehalten wird. Da die virtuellen Merkmale auch außerhalb des Bildbereiches liegen können, ist es ratsam, sie nur intern zu verwenden, und sie nicht für die Kamerapositionsschätzung, sondern nur für die weitere Merkmalsverfolgung einzusetzen.

Algorithmus 3 Einfacher Trifokalfilter

1. Berechne vorläufige Punktkorrespondenztripel über drei Bilder mittels symmetrischer Kosten aus Gleichung 49 und einem niedrigen Schwellwert $thres_{SIFT}$.
2. Berechne den Trifokaltensor mit GASAC und der Kostenfunktion aus Algorithmus 2.
3. Für alle Punktmerkmale aus Bild 1:
   (a) Sind Homographien bestimmt, berechne Guided Matching mittels Homographie für Bilder 1, 2 und 3 und gehe zum nächsten Merkmal.
   (b) Ansonsten bestimme die drei Fundamental- und Projektionsmatrizen aus dem Trifokaltensor.
   (c) Berechne für das Punktmerkmal mittels F_{12} die Epipolarregion in Bild 2 und bestimme die darin liegenden Kandidaten anhand des Sampsonabstands.
   (d) Berechne für jedes dieser Paare durch Triangulation aus P_1 und P_2 einen hypothetischen Raumpunkt und bestimme durch Rückprojektion in P_3 die vermutete Position in Bild 3.
      i. Prüfe, ob im Radius des durchschnittlichen Rückprojektionsfehlers um diese Position Merkmale bestimmt wurden und verwende diese Kandidaten.
      ii. Berechne paarweise die Kosten $cost_{SIFT}$ der SIFT-Dekriptoren für jedes Kandidatentripel nach Gleichung 48.
      iii. Verwirf jeden Kandidaten, der Kosten über dem Schwellwert 255 erzeugt.
   (e) Wähle das Kandidatentripel, das die geringsten Kosten verursacht.

10.3 Ergebnisse

Der vorgestellte trifokale Featuretracker kann wegen seiner drei Kameras nicht mit Standarddatensätzen verglichen werden, da diese mit nur einer Kamera aufgenommen wurden. Daher wird ein eigener Datensatz verwendet und zum Vergleich parallel mit dem Standard-KLT aus Abschnitt 7.1 mit und ohne affine Konsistenzprüfung analysiert. In [24] wurden Ergebnisse dieser Technik mit ungenaueren SIFT-Features und ohne die in Abschnitt 10.1.2 verbesserte Deskriptorenzuordnung vorgestellt. Dort ist bereits gezeigt, dass eine Farberweiterung des Standard-KLT nicht zur erhofften Verbesserung der Trackingergebnisse führt. Daher wurde er in dieser Arbeit nicht mehr berücksichtigt.

Für die Ergebnisevaluation wird eine 900 Triplets lange Sequenz entlang einer Häuserfront untersucht. Die Auflösung jedes einzelnen Bildes beträgt 2048×2048 Punkte bei 24bit Farbtiefe. Der Sequenzabschnitt wurde so gewählt, dass Start- und Endbild keinen gemeinsamen Überlappungsbereich haben und der Bildinhalt somit einmal komplett verschoben wurde.

Die Förstnerpunkte werden auf Farbbildern berechnet, jedoch erfolgt die darauf folgende SIFT-Deskriptorberechnung ausschließlich auf den Grauwerten des Bildes. Der KLT-Featuretracker arbeitet sowohl bei der Lokalisierung als auch bei der Beschreibung ausschließlich auf den Grauwertbildern. Zu beachten ist hier, dass bei den Ergebnissen des Trifokalfilters die Punktmerkmale bedingt durch den räumlichen Trifokalfilter nur im ge-

Kamera	Punkte pro Bild			Pfade	Pfadlänge	
	Mittel	Min	Max	Anzahl	Mittel	Max
KLT ohne affine Konsistenzprüfung						
C1	3446.3	732	4096	169576	15.3	637
C2	3427.5	764	4096	174182	14.9	629
C3	3433.6	500	4096	171209	15.2	630
KLT mit affiner Konsistenzprüfung						
C1	3344.5	625	4096	182931	13.5	637
C2	3334.1	455	4096	186764	13.2	630
C3	3349.6	611	4096	182271	13.6	630
Trifokales Featuretracking						
C1	1578.8	1095	1891	83842	19.6	638
C2	1569.5	1064	1865	83597	19.6	637
C3	1582.8	1121	1862	83759	19.7	637

Tabelle 3: Trackingergebnisse

meinsamen Überlappungsbereich der Bilder verfolgt werden können. Diese Flächen werden anhand der minimalen und maximalen Koordinaten der verfolgten Punktmerkmale in der jeweiligen Kamerasequenz ermittelt und ins Verhältnis zu den aufgespannten Flächen der KLT-Merkmale gesetzt. Sie betragen 86.9% für C1, 87.2% für C2 und 88.2% für C3.

Die Bewertung der Trackingergebnisse erfolgt statistisch über die Berechnung von minimaler, durchschnittlicher und maximaler Anzahl verfolgter Merkmale von Bild zu Bild. Zusätzlich wurden die Anzahl der Pfade, die mindestens drei Bilder lang waren, sowie maximale und durchschnittliche Pfadlänge für alle Punkte berechnet. Mit diesen Zahlen kann die Robustheit der verfolgten Punktmerkmale beurteilt werden.

Die Ergebnisse sind in Tabelle 3 aufgezeigt. Man sieht deutlich, dass die Verwendung von der affinen Kosistenzprüfung beim KLT zwar die Anzahl der Pfade steigert, jedoch die mittlere Pfadlänge und die Durchschnittsanzahl der verfolgten Punktmerkmale signifikant sinkt. Dies lässt den Schluss zu, das die affine Konsistenzprüfung häufiger gute Pfade zerreißt als die Robustheit steigert. Da die maximale Anzahl der verfolgten Punktmerkmale bei allen KLT-Pfaden gleich der Maximalanzahl der Punktmerkmale ist, gibt es in jedem Pfad mindestens drei aufeinander folgende Bilder, bei denen alle Punktmerkmale erfolgreich wiedergefunden wurden. Die durchschnittliche Anzahl der wiedergefundenen Punktmerkmale sowie die maximale Pfadlänge ist bei beiden KLT-Varianten sehr hoch.

Beim trifokalen Tracking sind im Vergleich zu den KLT-Varianten weniger als die Hälfte bei Pfadanzahl, der durchschnittlichen Anzahl und maximalen Anzahl von verfolgten Punktmerkmalen aufgezeichnet worden. Dennoch ist die minimale Anzahl von verfolgten Punkten und die durchschnittliche Pfadlänge erheblich höher als bei den KLT-Varianten. Auch die maximale Pfadlänge ist zumindest für die Kameras C2 und C3 geringfügig länger. Gerade die Pfadlänge und die minimale Anzahl von verfolgten Punkten sind jedoch für die Pfadrekonstruktion besonders wichtig, da diese Parameter die Qualität des späteren Bündelblockausgleiches signifikant verbessern. Auch liegen minimale, durchschnittliche und maximale Punktanzahl sehr viel näher beieinander, was auf eine geringere Streuung der Ergebnisse schließen lässt. Durch den Einsatz von virtuellen Punktmerkmalen, die für die Auswertung nicht berücksichtigt wurden, sind die unterschiedlichen Zahlen der drei Kameras beim trifokalen Tracking zu erklären. Der Hauptvorteil des trifokalen Trackings ist, dass die Merkmale synchron über mehrere Videoströme verfolgt werden, was mit der KLT-Technik nur unter großem Aufwand möglich ist. Durch diese Eigenschaft, die zunächst mehr Anforderungen an die Trackingergebnisse stellt, kann die Merkmalsverfolgung im Endeffekt sogar stabilisiert werden, da zwar weniger Pfade verfolgt werden, diese dafür jedoch länger und stabiler.

11 Kameraorientierung und Pfadextraktion

Die aneinandergereihten Kamerapositionen während einer Aufnahme bilden einen Pfad, der durch die hohe Bidfrequenz in der Regel kontinuierlich ist. Dennoch müsen bei der Extraktion des Kamerapfades aus Videotrackingdaten zahlreiche Sonderfälle berücksichtigt werden und die Kameraorientierungen müssen zusätzlich robust gegen leichte Verletzungen der Voraussetzungen sein. In der realen Welt können Videoaufnahmen selten unter idealen Laborbedingungen durchgeführt werden. Die grundsätzliche Voraussetzung, dass die Szene fest und unbeweglich ist, wird durch Verkehr, Fußgänger und Wind immer wieder veletzt. Diese Bewegungen werden bereits durch den Trifokalfilter aus Kapitel 10 gefiltert, weil sich das räumliche Model und das temporäre Model widersprechen. Dies gilt auch für spiegelnde Oberflächen, da sich die Spiegelungen abhängig von der Kameraposition verändern, sich aber nicht wie die übrigen Bildbereiche verhalten. Daher können auch sie bereits durch den Trifokalfilter aussortiert werden.

Die Positionen der Kameras werden in der klassischen Photogrammetrie über den räumlichen Rückwärtsschnitt bestimmt, also durch die Bestimmung von Bildkorrespondenzen zu bekannten 3D-Referenzpunkten im System. Dies erfolgt entweder mit bekannter und fixer intrinsischer Kalibrierung über [48] oder ohne Verwendung der intrinsischen Parameter über klassische Verfahren, wie sie in z.B. in [16] beschrieben sind. Diese Verfahren bergen jedoch den Nachteil, dass in jedem Bild genügend Referenzpunkte bekannt, sichtbar und erkennbar sein müssen. Des weiteren hängt die Qualität der Positionierung sehr stark mit der Genauigkeit der Referenzpunkte zusammen, und die Bestimmung solcher Referenzpunkte ist aufwändig und nicht immer möglich. Eine Alternative zur Kamerapositionsschätzung mittels räumlichem Rückwärtsschnitt ist die relative Orientierung ausschließlich aus Bildkorrespondenzen bei kalibrierten Kameras. Die theoretischen Grundlagen hierfür wurden bereits in Abschnitt 9.6 erläutert.

Im Straßenverkehr stoppt das System häufig, bei handgeführten Aufnahmen kommt es oft zu Drehungen um ein Projektionszentrum der Kamera oder im Bild sind nur Objekte zu sehen, die sehr weit entfert sind. In diesen Fällen schlägt vor allem die Translationsschätzung der Elementarmatrixberechnung fehl, da keine oder nur eine unzureichende Basislänge vorhanden ist. Auch ist bei der Aufnahme von planaren Objekten zu berücksichtigen, dass bei der Merkmalsverfolgung die Schätzung von temporalen Trifokalfiltern nicht möglich ist und nur Homographien bestimmt werden können. Daher ist es unbedingt erforderlich, die Elementarmatrix über Algorithmen zu bestimmen, die auch mit Korrespondenzen auf Ebenen funktionieren.

Diese Probleme können durch den Trifokalfilter allein nicht gelöst werden. In diesem Kapitel wird gezeigt, wie durch Triangulation über den Rahmen und über weitere Analyse der Trackingdaten die Problemfälle erkannt und kompensiert werden können. Da der Rahmen, auf dem die Kameras montiert sind, flexibel einstellbar ist und schon der Transport zu geringfügigen Änderungen an der Geometrie führt, ist die Orientierung über Bildkorrespondenzen direkt aus den aufgenommenen Bilddaten einer vorangegangenen oder nachträglich durchgeführten Kalibrierung mittels eines Kalibrierobjektes vorzuziehen. Eine klassische Kalibrierung ist nur für die Bestimmung der intrinsischen Parameter nötig. Daher wird zunächst gezeigt, wie die Orientierung der drei festen Kameras auf dem Rahmen bestimmt wird. Hierfür werden Bewertungskriterien für die Elementarmatrixberechnung untersucht.

Nachdem die Orientierung der Kameras auf dem Rahmen bestimmt ist, wird gezeigt, wie die Mehrdeutigkeiten bei der Elementarmatrixberechnung durch den Rahmen aufgelöst werden können und Techniken zur Ermittlung einer einheitlichen Skalierung entlang des Pfades vorgestellt. Zusätzlich wird das Gesammtverfahren zur Pfadextraktion um eine Stillstandschätzung erweitert, um nicht berechenbare Elementarmatrizen zu identifizieren.

11.1 Extraktion der richtigen Elementarmatrix

Die Bestimmung der Elementarmatrix E kann über das in [45] beschriebene Verfahren aus fünf Bildkorrespondenzen x_{ci} und x'_{ci}, $i\epsilon\{1,\cdots 5\}$ berechnet werden. Dabei wird die intrinsische Kalibrierung K der jeweiligen Kamera von den Bildkorrespondenzen abgezogen,

wodurch die zu bestimmenden Parameter der Kameramatrix P_2 auf eine Translation t und eine Rotation R reduziert werden können (vgl. Gleichung 53). Als Koordinatenreferenz für R und t dient die kanonische Kamera P_1(vgl. Gleichung 36).

$$\begin{aligned} x_{ci} &= K^{-1} x_i \\ x'_{ci} &= K'^{-1} x'_i \\ P_2 &= K^{-1} K \left[R \,|t\right] = \left[R \,|t\right] \end{aligned} \tag{53}$$

Allerdings liefert der verwendete Algorithmus bis zu zehn reelle, mathematisch korrekte Lösungen. Des weiteren kommt hinzu, dass im Normalfall zwar mehr als fünf Bildkorrespondenzen zur Verfügung stehen, diese jedoch ein meist unbekanntes Rauschen und auch zu einem nicht vernachlässigbaren Teil falsche Korrespondenzen beinhalten. In [54] ist gezeigt, dass eine Verwendung von mehr als den minimal benötigten fünf Punkten nicht zu einer Fehlerminimierung im Sinne der Kleinste-Quadrate-Lösung führt und es daher ratsam ist, auch bei mehr als fünf zur Verfügung stehenden Korrespondenzen nur fünf für die Beechnung zu verwenden. Da sich unter diesen fünf aber auch Ausreißer befinden oder die Punkte sich nicht für die Berechnung eignen können, weil sie kollinear oder auf einer kritischen Oberfläche sind, muss analysiert werden, ob sich unter den Ergebnissen überhaupt eine richtige Lösung befindet. Erst dann kann versucht werden, aus den Ergebnissen die richtige Lösung zu extrahieren.

Die robuste Bestimmung einer geeigneten Teilmenge kann wieder über GASAC erfolgen. Dazu wird ein robustes Fehlermaß benötigt, das bei jedem der zufällig bestimmten Kandidatensätze die gültigen Lösungen findet, sie einheitlich bewertet und die beste auswählt. Gültige Lösungen müssen bestimmte eigenschaften erfüllen, die im Folgenden erläutert werden.

Aus jeder der berechneten Elementarmatrizen $E_{[1\ldots10]}$ können eine Rotationsmatrix R und ein Translationsvektor t der Länge Eins bestimmt werden [22]. Aus diesen beiden Ergebnissen sind wegen der ungerichteten Strahlengeometrie vier Kombinationen für die Kameramatrix P_2 möglich:

$$\begin{aligned} P_a &= \left[R \,|{+}t\right] \qquad & P_b &= \left[R^{-1} \,|{+}t\right] \\ P_c &= \left[R \,|{-}t\right] \qquad & P_d &= \left[R^{-1} \,|{-}t\right] \end{aligned} \tag{54}$$

Die Kandidaten für die richtige Lösung können somit eingeschränkt werden, indem die fünf Punkte trianguliert werden und nur die Lösungen in Betracht kommen, bei denen alle Punkte vor beiden Kameras liegen und somit überhaupt sichtbar sind. Diese Bedingung muss auch auf alle weiteren Punktkorrespondenzen anwendbar sein, die nicht zur Bestimmung verwendet wurden. In der Literatur wird diese physikalische Bedingung *'Cherality-constraint'* [46] oder *'Chirality-constraint'* [20] genannt. Die Bedingung fordert, dass Raumpunkte X vor Kamera P mit Projektionszentrum C liegen, wenn gilt:

$$\begin{gathered} w = P^{31}\left(X_x - C_x\right) + P^{32}\left(X_y - C_y\right) + P^{33}\left(X_z - C_z\right) \\ w > 0 \end{gathered} \tag{55}$$

Diese Bedingung schließt jedoch nicht alle falschen Lösungen aus, auch wenn mehr als fünf Punkte überprüft werden [54]. In [66] ist sogar aufgezeigt, dass es unter bestimmten Bedingungen zwei mathematisch nicht zu unterscheidene Lösungen gibt, weil die Kamera sich schräg entlang einer parallelen Ebene bewegt hat. Ferner ist die Triangulation der Bildkorrespondenzen rechenaufwändig, was die Eignung dieses Tests bei sehr vielen Kamerapositionsschätzungen beeinträchtigt. Daher wird der *cherality check* nur auf den fünf zur Berechnung verwendeten Punkte ausgeführt.

Da die Eigenschaften der Fundamentalmatrix auch für die Elementarmatrix gelten (vgl. Abschnitt 9.4 und 9.6), beschreibt der Abstand der korrespondierenden Punkte x_c und x'_c zu ihrer Epipolargeraden ein Fehlermaß, um die richtige Lösung aus der Lösungsmenge zu extrahieren. Die Summe der Abstände wird in der Literatur als Sampsonabstand oder -fehler bezeichnet und berechnet sich aus:

$$l = Ex_c \qquad l' = E^T x'_c$$
$$err_{sampson} = \sum_i \frac{(x'_c E x_c)^2}{{l_x}^2 + {l_y}^2 + {l'_x}^2 + {l'_y}^2} \tag{56}$$

Da dieses Fehlermaß im Vergleich zur Triangulation schnell zu berechnen ist, kann man es auf eine große Anzahl von Punktkorrespondenzen anwenden. Der Sampsonabstand ist für die fünf initialen Punkte immer exakt Null. Wegen der nicht perfekten Lokalisation der zusätzlich verfügbaren Punkte entstehen jedoch bei diesen Punkten Fehler, die aber für die richtige Lösung sehr gering sein müssen. Allerdings ist in [54] gezeigt, dass die Elementarmatrix mit dem geringsten Sampsonabstand bei sehr geringem Rauschen auch die richtige Lösung ist. Um die richtige Lösung auch bei höherem Rauschanteil zuverlässig bestimmen zu können, müssen daher zusätzliche geometrische Eigenschaften hinzugezogen werden.

11.2 Orientierung mehrerer Kameras auf einem Rahmen

Das Problem der Lösungsidentifikation kann auf zwei Arten angegangen werden: Entweder man weiß, wie die Kameras ungefähr zueinander standen, und filtert inkorrekte Lösungen manuell, oder man versucht, zusätzliche, möglichst allgemein gültige Bedingungen einzuführen. Für eine vollautomatische Analyse kommt nur Letzteres in Betracht. Es wird daher angenommen, dass die beiden Kamerapaare (P_1, P_2) und (P_1, P_3) eine recht ähnliche Skalierung und ein teilweise gemeinsames Sichtfeld haben und dass die Basis in einem guten Verhältnis zum Objektabstand steht. Aus diesen Annahmen folgt, dass beide Kameras dieselben Objektpunkte triangulieren können und die 3D-Punkte hinreichend genau bestimmt sind. So kann über die Ähnlichkeit der triangulierten 3D-Punkte aus den zwei Kamerapaaren auf die richtige Lösung geschlossen werden.

Algorithmus 4 Relative Orientierung für drei Kameras

1. Führe eine GASAC zur Berechnung eines geeigneten Punktesatzes aus. Das Gütekriterium berechnet sich aus:
   (a) Extrahiere die vier P-Matrizen (Gleichung 54) und prüfe, ob mindestens eine davon die zur Berechnung verwendeten fünf Punkte vor den Kameras hat.
   (b) Ordne die verbleibenden Lösungen nach ihrem Sampsonabstand (Gleichung 56).
2. Filtere die Ausreißer mit 3σ des Sampsonabstandes.
3. Trianguliere die guten Punkte der Lösungen und prüfe, ob alle Punkte vor den Kameras liegen. Wenn nicht all Punkte vor den Kameras liegen, verwirf die Lösung.
4. Berechne für alle Lösungen den räumlichen Abstand der Punkte zur Referenzkamera.
5. Berechne das zweite Kamerapaar mit denselben Kriterien.
6. Berechne den Maßstabsfaktor sämtlicher Permutationen der Lösungen nach Gleichung 76.
7. Wähle das Lösungspaar mit dem geringsten Restfehler nach Gleichung 57.
8. Führe einen Bündelausgleich für das so bestimmte Kameratripel durch.

Zuerst werden zwei Punktwolken $X_{1,2}$ und $X_{1,3}$ der beiden Kamerapaare trianguliert. Da sich beide Punktwolken auf dieselbe Referenzkamera P_1 beziehen, ist abgesehen von der Skalierung eine Transformation der Kameras nach Abschnitt 11.3.2 nicht nötig. Nachdem ein gemeinsamer Skalierungsfaktor der zwei Punktwolken mittels der Technik aus Abschnitt 11.3.3 bestimmt wurde, wird durch Berechnung des Restfehlers geprüft, wie gut diese Punktwolken zueinander passen:

$$err_{tri} = \sum_i |X_{1,2} - \mu X_{1,3}| \tag{57}$$

Die Lösung mit dem geringsten Restfehler wird als korrekt für die beiden Kamerapaare angenommen. Allerdings ist dieser Test rechentechnisch aufwändig, weil zusätzlich zur Triangulation noch eine Abstandsmessung und eine SVD durchgeführt werden müssen. Des weiteren müsste er für jede mögliche Kombination der zwei Elementarmatrixberechnungen durchgeführt werden, was im Extremfall 100 Kombinationen sein können. Um frühzeitig ungültige Lösungen zu vermeiden, werden die Berechnungsschritte aus Algorithmus 4 durchgeführt.

11.3 Kameraorientierung entlang des Aufnahmepfades

Eine Orientierung der Kameras nach Abschnitt 11.2 ist für Tausende von Bildern einer Videosequenz aus mehreren Gründen nicht praktikabel: Die Kameras bewegen sich von einem Tripel zum Nächsten teilweise nur sehr wenig, daher ist eine ausreichende Basislänge für eine robuste Triangulation nicht immer gegeben (vgl. Algorithmus 4 Schritt 4). Des weiteren müssen die unterschiedlichen Skalierungsfaktoren für die Translationslängen robust bestimmt werden, um ein einheitliches Koordinatensystem für alle Kameras und damit auch alle triangulierten Punkte zu ermöglichen, was bei Basislängen von wenigen Zentimetern ungenau ist.

Da alle Bildtripel von einem festen Rahmen aus aufgenommen werden, dient dieser starre Rahmen als globale Skalierung und bestimmt den globalen Maßstab. Sämtliche zeitlichen Skalierungen müssen sich auf diesen Maßstab beziehen. Die geometrischen Bedingungen, die dadurch an die Bewegung der Kameras gestellt werden, helfen auch bei der Extraktion der richtigen Lösung aus der Elementarmatrixberechnung.

11.3.1 Geometrische Bedingungen durch den Rahmen

Der feste Rahmen koppelt die Bewegungen der Kameras aneinander, wodurch aus der Bewegung einer Kamera von einem Bild zum nächsten auf die der anderen Kameras geschlossen werden kann (vgl. Abbildung 14). Die Kamerapositionen sind durch P bestimmt, die Transformationsmatrizen zwischen zwei Kamerapositionen werden mit T gekennzeichnet, wobei der untere Index die Nummer der Kamera angibt, der obere Index die Nummer des Bildes. Die Transformation ΔT_{12} repräsentiert die relative Orientierung der Kamera P_2 zur Kamera P_1 auf dem Rahmen. Die Länge der Basis ΔT_{12} ist die Referenzskala des Systems und wird für die folgenden Berechnungen auf Eins gesetzt.

Zunächst werden über die Berechnung der Elementarmatrix die auf Eins normierten Transformationsschätzungen $\widehat{T}_1^2$ und $\widehat{T}_2^2$ bestimmt. Bewegt sich nun die Kamera P_1^1 durch die Transformation $T_1^2 = \lambda_1 \widehat{T}_1^2$ auf die Position P_1^2, kann die Transformation $T_2^2 = \lambda_2 \widehat{T}_2^2$ aus der Position P_1^2 und der Transformation ΔT_2 ermittelt werden. Weicht nun die Schätzung der Elementarmatrix $\widehat{T}_2^2$ von der mittels T_1^2 und ΔT_2 berechneten Position ab, ist eine der gewählten Lösungen für die Elementarmatrizen , die für die Berechnung von $\widehat{T}_1^2$ und $\widehat{T}_2^2$ verwendet wurden, nicht korrekt. Für die gekoppelte Bewegung auf dem Rahmen gilt:

$$\Delta T_{12} T_1^2 = T_2^2 \triangle T_{12} \tag{58}$$

Das Problem dieser Bedingung liegt in den unbekannten Skalierungsfaktoren λ_1 und λ_2. Die Transformationen werden daher in einen (3×3) −Rotations-, einen (1×3) −Translations- und einen Skalierungsanteil aufgespalten:

$$T_b^a = \begin{bmatrix} & R_b^a & & \lambda_b t_b^a \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{59}$$

Da die Skalierung bei der Rotation nicht von Bedeutung ist, gilt für die Rotation:

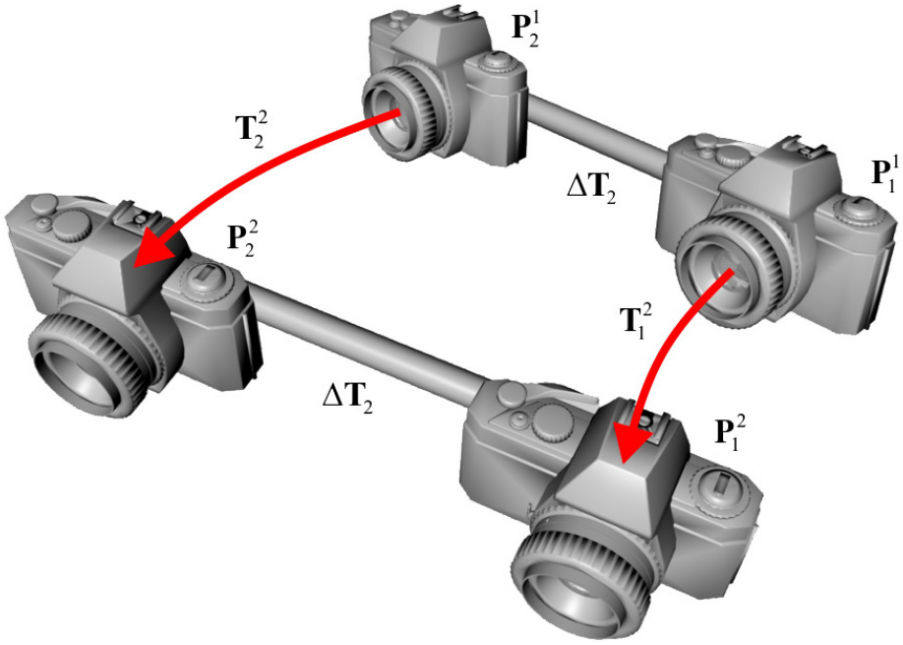


Abbildung 14: Bewegung der Kameras auf dem Rahmen

$$\Delta R_{12} R_1^2 = R_2^2 \triangle R_{12} \quad \Leftrightarrow \quad \triangle R_{12}^T R_2^{2T} \Delta R_{12} R_1^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{60}$$

Die Abweichung der sequenziellen Rotationen von der Identitätsmatrix bestimmt das Fehlermaß. Diese Abweichung kann durch drei Eulerrotationswinkel beschrieben werden. Der größte Absolutwert dieser Winkel definiert eine obere Schranke des Rotationsfehlers und wird als Fehlermaß verwendet:

$$\begin{bmatrix} \phi \\ \kappa \\ \omega \end{bmatrix} = euler(\triangle R_{12}^T R_2^{2T} \Delta R_{12} R_1^2) \tag{61}$$

$$err_{rot} = max\left(|\phi|, |\kappa|, |\omega|\right)$$

Bei der Translation müssen die Skalierungsfaktoren berücksichtigt werden, was die Bedingungen für die Translation etwas komplizierter gestaltet:

$$R_1^2 \Delta t_{12} + \lambda_1 t_1^2 = \lambda_2 \Delta R_{12} t_2^2 + \Delta t_{12} \tag{62}$$

Diese Gleichung lässt sich in ein lineares Gleichungssystem transformieren, welches mittels SVD gelöst werden kann:

$$\begin{bmatrix} t_1^2 & -\Delta R_{12} t_2^2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \Delta t_{12} - R_1^2 \Delta t_{12} \tag{63}$$

Da das Gleichungssystem 63 überbestimmt ist, kann über die Residuen ein Restfehlervektor r bestimmt werden:

$$r = \begin{bmatrix} t_1^2 & -\Delta R_{12} t_2^2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} - \left(\Delta t_{12} - R_1^2 \Delta t_{12}\right) \tag{64}$$

Die euklidische Länge von r wird als Fehlermaß der Translation verwendet:

$$err_{trans} = \sqrt{(r^x)^2 + (r^y)^2 + (r^z)^2} \tag{65}$$

Da das System für drei Kameras ausgelegt ist, werden die Gleichungen 60 und 63 jeweils um eine Kamera erweitert. Für die Rotation ergeben sich zwei weitere Fehlermaße, die aus diesen Gleichungen bestimmt werden können:

$$\begin{aligned} \Delta R_{13} R_1^2 = R_3^2 \triangle R_{13} \Leftrightarrow \triangle R_{13}^T R_3^{2T} \Delta R_{13} R_1^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ \Delta R_{23} R_2^2 = R_3^2 \triangle R_{23} \Leftrightarrow \triangle R_{23}^T R_3^{2T} \Delta R_{23} R_2^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{66}$$

Bei der Translation ist es wichtig, die Parameter symmetrisch zu bestimmen, damit jede Transformation gleich gewichtet ist. Jeder Weg wird also einmal im und einmal gegen den Uhrzeigersinn gegangen. Dabei müssen einige neue Transformationen auf dem Rahmen eingeführt werden, die sich aus den vorhandenen Transformationen ΔT_1, ΔT_2 und ΔT_3 errechnen lassen. Die Matrix ΔT_{ab} und die entsprechende Rotation ΔR_{ab} bzw. Translation Δt_{ab} beschreiben die Transformation der Kamera P_a^1 auf die Kamera P_b^1:

$$\Delta T_{ab} = \Delta T_a^{-1} \cdot \Delta T_b \tag{67}$$

Daraus ergeben sich neue Gleichungssysteme:

$$\begin{aligned} R_1^2 \Delta t_{13} + \lambda_1 t_1^2 &= \lambda_3 \Delta R_{13} t_3^2 + \Delta t_{13} \\ R_2^2 \Delta t_{23} + \lambda_2 t_2^2 &= \lambda_3 \Delta R_{23} t_3^2 + \Delta t_{23} \\ R_2^2 \Delta t_{21} + \lambda_2 t_2^2 &= \lambda_1 \Delta R_{21} t_1^2 + \Delta t_{21} \\ R_3^2 \Delta t_{32} + \lambda_3 t_3^2 &= \lambda_2 \Delta R_{32} t_2^2 + \Delta t_{32} \\ R_3^2 \Delta t_{31} + \lambda_3 t_3^2 &= \lambda_1 \Delta R_{31} t_1^2 + \Delta t_{31} \end{aligned} \tag{68}$$

Das lineare Gleichungssystem für die Bestimmung der Skalierungsfaktoren wird ebenfalls erweitert:

$$\begin{bmatrix} t_1^2 & -\Delta R_{12} t_2^2 & 0 \\ t_1^2 & 0 & -\Delta R_{13} t_3^2 \\ 0 & t_2^2 & -\Delta R_{23} t_3^2 \\ -\Delta R_{21} t_1^2 & t_2^2 & 0 \\ 0 & -\Delta R_{32} t_2^2 & t_3^2 \\ -\Delta R_{31} t_1^2 & 0 & t_3^2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} \Delta t_{12} - R_1^2 \Delta t_{12} \\ \Delta t_{13} - R_1^2 \Delta t_{13} \\ \Delta t_{23} - R_2^2 \Delta t_{23} \\ \Delta t_{21} - R_2^2 \Delta t_{21} \\ \Delta t_{32} - R_3^2 \Delta t_{32} \\ \Delta t_{31} - R_3^2 \Delta t_{31} \end{bmatrix} \tag{69}$$

Zur Fehlerüberprüfung werden hier die Residuen für die sechs Translationen paarweise berechnet und das Maximum dieser drei Translationsfehler als obere Schranke verwendet:

$$
\begin{aligned}
err_{1-2} = & \left(\left| \begin{bmatrix} t_1^2 & -\Delta R_{12} t_2^2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} - \left(\Delta t_{12} - R_1^2 \Delta t_{12} \right) \right| \right. \\
& \left. + \left| \begin{bmatrix} -\Delta R_{21} t_1^2 & t_2^2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} - \left(\Delta t_{21} - R_2^2 \Delta t_{21} \right) \right| \right) \\
err_{1-3} = & \left(\left| \begin{bmatrix} t_1^2 & -\Delta R_{13} t_3^2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_3 \end{bmatrix} - \left(\Delta t_{13} - R_1^2 \Delta t_{13} \right) \right| \right. \\
& \left. + \left| \begin{bmatrix} -\Delta R_{31} t_1^2 & t_3^2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_3 \end{bmatrix} - \left(\Delta t_{31} - R_3^2 \Delta t_{31} \right) \right| \right) \\
err_{2-3} = & \left(\left| \begin{bmatrix} t_2^2 & -\Delta R_{23} t_3^2 \end{bmatrix} \begin{bmatrix} \lambda_2 \\ \lambda_3 \end{bmatrix} - \left(\Delta t_{23} - R_2^2 \Delta t_{23} \right) \right| \right. \\
& \left. + \left| \begin{bmatrix} -\Delta R_{32} t_2^2 & t_3^2 \end{bmatrix} \begin{bmatrix} \lambda_2 \\ \lambda_3 \end{bmatrix} - \left(\Delta t_{32} - R_3^2 \Delta t_{32} \right) \right| \right)
\end{aligned}
\tag{70}
$$

$$
err_{trans} = max\left(err_{1-2}, err_{1-3}, err_{2-3} \right)
$$

Der Rotationsfehler err_{rot} ist in Grad angegeben und kann gegen einen Schwellwert geprüft werden, wobei sich eine Winkeltoleranz von 0.5° als sinnvoller Schwellwert erwiesen hat. Der Translationsfehler err_{trans} gibt die Länge der maximalen Abweichung an und bezieht sich auf die Basislänge Eins. Ein sinnvoller Schwellwert ist hier 1% der Basislänge.

Diese geometrischen Bedingungen lassen sich auf beliebig viele Kameras auf dem Rahmen erweitern. Zu beachten ist, dass für die Bedingungen nur die Orientierung der Kameras auf dem Rahmen benötigt werden und keine gemeinsamen 3D-Punkte. Diese Technik kann daher auch für Kamerasysteme verwendet werden, deren Sichtfeld keine gemeinsame Schnittmenge aufweist, z.B. vorwärts und rückwärts ausgerichtete Kameras auf einem Fahrzeug.

Die λ-Parameter sind nicht immer geeignet, um die Skalierungsanpassung der drei Kameras zu bestimmen. Werden die Kameras nicht rotiert, sind die drei Matrizen R_i^2 die Iditentitätsmatrizen. Dadurch wird offensichtlich die rechte Seite der Gleichungen 63 bzw. 69 Null und die Gleichungen werden zu einem homogenen Gleichungssystem der Form $A\left[\lambda_1, \lambda_2, \lambda_3\right]^T = 0$. Die drei Skalierungsfaktoren definieren somit einen Nullvektor, der bis auf einen gemeinsamen Skalierungsfaktor unbestimmt ist. Dadurch ist der reale Zahlenwert dieser drei Parameter unbestimmt und kann sogar Null betragen, nämlich dann, wenn das Kamerasystem gar nicht bewegt wurde.

11.3.2 Transformation in ein gemeinsames Koordinatensystem

Der 5-Punkte-Algorithmus geht von einer Translation der Länge Eins aus. Sowohl die beiden Kamerapaare auf dem Rahmen als auch die zwei Kamerapaare aus drei aufeinander folgenden Bildern sind jedoch in der Regel unterschiedlich skaliert. Die korrekte Bestimmung eines globalen Maßstab muss daher nachträglich erfolgen. Hierbei ist zu beachten, dass der Maßstabsfaktor Null ist, wenn die Kameras nicht bewegt wurden, was zu zusätzlichen Problemen führen kann. Grundsätzlich besteht die Aufgabe darin, das Kamerapaar (P_1, P_2) an eine bestehende Kette von Kameras P_n anzufügen, wobei davon ausgegangen wird, dass die Kamera P_1 mit der Kamera P_n identisch ist und geometrisch so angepasst werden muss, dass beide dasselbe Koordinatensystem verwenden. Dazu gibt es zwei Verfahren:

1. Transformation der Raumpunkte in ein gemeinsames Koordinatensystem
2. Transformation der Kameras auf die gemeinsame Referenzkamera

Mit dem ersten Verfahren soll eine Homographie H_{coord} bestimmt werden, um die Raumpunkte X' aus der Triangulation von (P_1, P_2) auf die bisherigen Raumpunkte X zu transformieren:

$$X = H_{coord} X' \tag{71}$$

Dabei werden Rotation, Translation und Skalierung gleichzeitig über eine SVD bestimmt. Allerdings beträgt die Basislänge zweier aufeinander folgendener Bilder in einer Bildsequenz oft nur wenige Zentimeter, während die aufgenommenen Objekte einige Meter voneinander entfernt sind. Durch dieses ungünstige Verhältnis von Basis zu Objektabstand ist die Triangulation aus einem Kamerapaar so schwach bestimmt, dass die direkte Transformation der Raumpunkte zu ungenau ist. Die Triangulation erweist sich als gänzlich unmöglich, wenn sich die Kamera P_2 zur Referenzkamera nicht bewegt hat oder lediglich rotiert wurde.

Das zweite Verfahren besteht darin, die Kamera P_2 direkt an die Kamera P_n zu transformieren, indem die Referenzkamera P_1 mittels einer Homographie auf P_n transformiert wird:

$$\begin{aligned} T_0 &= \begin{bmatrix} & P_1 & \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ T_I &= \begin{bmatrix} & P_n & \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ H_{cam} &= T_I^{-1} T_0 \\ P_{n+1} &= P_2 H_{cam}^{-1} \end{aligned} \tag{72}$$

Da P_1 die kanonische Kamera und T_0 daher die Identitätsmatrix ist, vereinfacht sich die Gleichung zu:

$$P_{n+1} = P_2 \begin{bmatrix} & P_n & \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{73}$$

Allerdings ist hierbei die Skalierung der Kamerabasis von P_2 nicht bestimmt. Die Anpassung an einen globalen Skalierungsfaktor muss daher vor der Transformation von Gleichung 7 erfolgen:

$$P_2 = [R \,|\mu_n t] \tag{74}$$

11.3.3 Bestimmung des Maßstabs

Da die Basislänge des neuen Kamerapaares unbekannt ist, muss ein Skalierungsparameter μ_n bestimmt werden, damit die korrekt skalierten Raumpunkte folgende Gleichung erfüllen:

$$X = \mu_n H_{cam} X' \tag{75}$$

Die Qualität dieser Skalierung hat einen starken Einfluss auf die Qualität der Rekonstruktion. Daher muss die Bestimmung sehr robust sein und sollte nicht nur aus einem Raumpunkt berechnet werden, was theoretisch möglich wäre. Zur Bestimmung des Skalierungsfaktors werden die folgenden drei Möglichkeiten untersucht:

1. Bestimmung über den Abstand zur Kamera: Die Skalaanpassung kann über die Abstände der Raumpunkte X_i zum Projektionszentrum C_n der Kamera P_n und den Abständen der Raumpunkte X_i' zum Projektionszentrum C_1 der Referenzkamera P_1 definiert werden:

$$|C_n - X_i| = \mu_n \, |C_1 - X_i'| \tag{76}$$

Hierbei sei erneut angemerkt, dass P_1 die kanonische Kamera ist und daher Gleichung 76 vereinfacht werden kann:

$$|C_n - X_i| = \mu_n \, |X_i'| \tag{77}$$

Eine Lösung für den Skalierungsfaktor μ_n kann für alle Raumpunkte über eine SVD berechnet werden. Da die SVD jedoch alle Abstände gleich behandelt, die Genauigkeit der Triangulation aber mit Zunahme der Tiefe abnimmt, gewichtet sie den Fehler bei entfernten Punkten im Vergleich zu nahe liegenden Punkten zu stark. Um diesem Effekt entgegenzuwirken, sollten nur die Punktpaare für die Skalierungsanpassung verwendet werden, deren Abstand $|C_n - X_i|$ einen gewissen Wert nicht überschreitet.

2. Bestimmung über die Tiefe zur Kamera: Da die Triangulation entfernter Punkte größere Fehler entlang der Sichtstrahlen erzeugt, wird der Abstand auf diesem Strahl verwendet, um die Skala anzupassen. Die Länge des Sichtstrahls vom Raumpunkt zu einer normierten Kamera P mit Projektionszentrum C berechnet sich aus:

$$depth\,(X_i, P, C) = P^{31}(X_i^x - C^x) + P^{32}(X_i^y - C^y) + P^{33}(X_i^z - C^z) \tag{78}$$

Der Skalierungsfaktor berechnet sich aus der Lösung dieser Gleichung mittels SVD:

$$depth\,(X_i, P_n, C_n) = \mu_n \cdot depth\,(X_i', P_1, C_1) \tag{79}$$

Wie schon im zweiten Verfahren beschrieben wird der Abstand zur Kamera P_n gegen einen Schwellwert geprüft, zu weit entfernte Punkte werden verworfen.

3. Bestimmung über Rückprojektion: Befinden sich die Kameras P_n und P_1 im selben Koordinatensystem, ist von Kamera P_2 nur noch die Skalierung der Translation unbekannt (Gleichung 74). Daher muss für die projizierten Punkte $x_{2,i}$ gelten:

$$x_{2,i} = P_2 X_i = [R\,|\mu_n t]\, X_i \tag{80}$$

Da die Bildkoordinaten normiert sind, muss folgendes Gleichungssystem gelöst werden:

$$w_j = P_2^{31} X_j^x + P_2^{32} X_j^y + P_2^{33} X_j^z$$

$$\begin{bmatrix} P_2^{14} - P_2^{34} x_{2,0}^x \\ P_2^{24} - P_2^{34} x_{2,0}^y \\ \vdots \\ P_2^{14} - P_2^{34} x_{2,i}^x \\ P_2^{24} - P_2^{34} x_{2,i}^y \end{bmatrix} \mu_n = \begin{bmatrix} x_{2,0}^x \cdot w_0 - (P_2^{11} X_0^x + P_2^{12} X_0^y + P_2^{13} X_0^z) \\ x_{2,0}^y \cdot w_0 - (P_2^{21} X_0^x + P_2^{22} X_0^y + P_2^{23} X_0^z) \\ \vdots \\ x_{2,i}^x \cdot w_i - (P_2^{11} X_i^x + P_2^{12} X_i^y + P_2^{13} X_i^z) \\ x_{2,i}^y \cdot w_i - (P_2^{21} X_i^x + P_2^{22} X_i^y + P_2^{23} X_i^z) \end{bmatrix} \tag{81}$$

Auch hier werden Punkte verworfen, die zu weit entfernt sind.

Diese Verfahren wurden empirisch an realen Daten getestet. Dabei wurde die Hälfte der Daten zur Bestimmung des Skalierungsfaktors verwendet und die andere Hälfte zur Berechnung des Fehlers. Das Fehlermaß wurde über den Rückprojektionsabstand der skalierten Raumpunkte zu den Referenzpunkten in der jeweiligen Kamera berechnet. Da von den Kameras die intrinsische Kalibrierung abgezogen wurde, sind die Rückprojektionsfehler zahlenmäßig sehr klein und es wurden Minimum, Maximum und Durchschnittswert mit 10^7 multipliziert. Von 183 Kamerabewegungen wurden minimaler, maximaler und durchschnittlicher Fehler analysiert. Zusätzlich wurde gezählt, welche Methode im Vergleich zu den beiden anderen den niedrigsten und welche den höchsten Fehler aufwies. Zuletzt wurde der durchschnittliche Fehler über die Messwerte berechnet, bei denen die Methode im Vergleich zu den anderen das schlechteste Ergebnis produzierte, um abzuschätzen, ob es sich hierbei generell um Ausreißer handelt. Die Ergebnisse sind in Tabelle 4 zu sehen. Zu bemerken ist, dass der Fehler der Abstandsmethode in einem Fall exakt dem der Tiefenmethode entsprach, weshalb die niedrigsten Ergebnisse in der Summe nicht 183 ergeben.

Aus diesen Messwerten ist ersichtlich, dass die Rückprojektionsmethode eindeutig zu den robustesten Ergebnissen führt. Bei Verwendung der anderen Methoden sind die maximalen Fehler fast doppelt so hoch, ferner produziert die Rückprojektionsmethode am häufigsten

Fehler	Abstand	Tiefe	Rückprojektion
Min	3.3176	3.3173	3.3136
Max	110.1733	112.5690	55.1833
Ø	11.0952	11.0975	10.0074
Niedrigster	15	24	143
Höchster	99	46	38
Ø Höchster	11.8661	10.4766	9.8359

Tabelle 4: Analyse Skalaanpassung

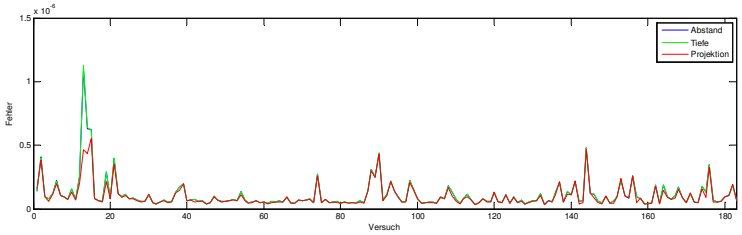


Abbildung 15: Rückprojektionsfehler

das beste Ergebnis. An den Stellen, an denen diese Methode verglichen mit den anderen zu schlechteren Ergebnissen führt, sind die Ergebnisse im Vergleich zu allen übrigen Messungen überdurchschnittlich gut, was bedeutet, dass an diesen Stellen der Fehler insgesamt bei allen Methoden sehr gering war. Daraus folgt im Umkehrschluss, dass diese Methode besonders bei kritischen Konfigurationen mit überdurchschnittlichem Fehler im Vergleich zu den anderen Methoden das beste Ergebnis liefert. Die beiden anderen Verfahren ergeben sehr ähnliche Werte. Auch wenn die Tiefenmethode den maximalen absoluten Fehler aufweißt, erzeugt sie seltener die im Vergleich höchsten Fehler und auch der Durchschnittswert der hohen Fehler ist niedriger als der Gesamtdurchschnitt dieser Methode. Die Fehler sind in Abbildung 15 aufgezeigt. Die blaue Kurve ist fast immer durch die grüne verdeckt, was zeigt, wie ähnlich sich Tiefen- und Abstandsmethode sind.

Zu beachten ist, dass die Triangulation der Referenzpunkte für die Rückprojektionstechnik auch über den Rahmen erfolgen kann. Die Punkte können dann über die im Vergleich zur zeitlichen Basis sehr viel längere Rahmenbasis trianguliert werden, zusätzlich ist es möglich, dass die Kamera gar nicht bewegt wird. In diesem Fall wird beim Rückprojektionsansatz die Skalierung auf Null gesetzt. Die beiden anderen Verfahren benötigen auf jeden Fall Rauminformation aus der Bewegung; bei geringen Bewegungen kann es sein, dass diese nicht ausreicht.

11.3.4 Stillstandsschätzung

Auch wenn die Rückprojektionstechnik still stehende Kameras korrekt skalieren kann, tritt bei der relativen Orientierung still stehender Kameras ein Problem auf. Der *cherality check* aus Abschnitt 11.1 setzt voraus, dass die fünf Punkte für die Orientierung trianguliert werden können und sich im Sichtfeld und somit vor den beiden Kameras befinden. Um die Triangulierung durchführen zu können, muss daher herausgefunden werden, ob sich die Kameras überhaupt bewegt haben.

In [47] wird gesagt, dass mindestens eine Bewegung der Punktkorrespondenzen von 10% der Bildgrößen erforderlich ist, um die Elementarmatrix zu berechnen. Allerdings kann dadurch nicht erkannt werden, ob sich die Kamera nur gedreht hat. Dieser Fall führt zu starken Bildbewegungen, die jedoch nicht aus Translation der Kamera resultieren.

In der Arbeit von [64] wird eine Technik vorgeschlagen, um zu prüfen, ob überhaupt eine Translation vorliegt. Dazu wird der Rotationsanteil R der Elementarmatrix ermittelt und es wird geprüft, ob eine Rotation des Bildes ausreicht, um die kalibrierten Punkte von einem Bild ins andere zu transferieren. Dies kann durch einen einfachen Test für alle Punktepaare

geprüft werden. Hierbei muss folgender Fehlerwert für alle Punkte kleiner als der Schwellwert $thresh_{motion}$ sein:

$$\frac{|x_i' \times Rx_i|}{|x_i|\,|x_i'|} < thresh_{motion} \tag{82}$$

Diese Gleichung prüft den eingeschlossenen Winkel der beiden Punktvektoren. Ein geeigneter Schwellwert lässt sich aus einer erlaubten Translationstoleranz zweier Punkte in einer Bildecke, die weit vom Bildhauptpunkt entfernt ist, errechnen. Da der Bildhauptpunkt meist in der Bildmitte liegt, können die zu testenden Bildpunkte der Einfachheit halber in die obere linke Bildecke gelegt werden. Die Translationstoleranz, bei der noch keine signifikante Kamerabewegung zu erkennen ist, wird hier mit zwei Bildpunkten angegeben. Der Schwellwert für den Bildhauptpunkt (b^x, b^y) errechnet sich aus:

$$thres_{motion} = min\left(\frac{\left|\left|\begin{bmatrix} -b^x \\ -b^y \\ 1 \end{bmatrix} \times \begin{bmatrix} 2-b^x \\ -b^y \\ 1 \end{bmatrix}\right|\right|}{\left|\left|\begin{bmatrix} -b^x \\ -b^y \\ 1 \end{bmatrix}\right|\right|\left|\left|\begin{bmatrix} 2-b^x \\ -b^y \\ 1 \end{bmatrix}\right|\right|}, \frac{\left|\left|\begin{bmatrix} -b^x \\ -b^y \\ 1 \end{bmatrix} \times \begin{bmatrix} -b^x \\ 2-b^y \\ 1 \end{bmatrix}\right|\right|}{\left|\left|\begin{bmatrix} -b^x \\ -b^y \\ 1 \end{bmatrix}\right|\right|\left|\left|\begin{bmatrix} -b^x \\ 2-b^y \\ 1 \end{bmatrix}\right|\right|}\right) \tag{83}$$

Diese Prüfung wird vor dem *cherality check* für alle Punkte durchgeführt. Zeigt sich, dass ein Bildpaar keine ausreichende Translation aufweist, wird der Translationsvektor, der aus der Elementarmatrix extrahiert wird, auf Null gesetzt und es kann nur der Sampsonabstand als Gütekriterium für diese Elementarmatrix verwendet werden.

11.4 Bestimmung des Kamerazentrums aus der Projektionsmatrix

Liegen die Projektionsmatrizen vor, werden speziell für den Bündelausgleich (vgl. Abschnitt 11.5) die Projektionszentren benötigt, die durch die Projektionsmatrix pararametrisiert sind. Generell gilt für das Projektionszentrum C einer Projektionsmatrix P:

$$P \cdot C = \vec{\mathbf{0}} \tag{84}$$

Das Projektionszentrum ist demnach ein rechter Nullvektor der Matrix P. Die Berechnung dieses Vektors kann über drei Verfahren erfolgen, die im Anschluss kurz erläutert werden.

Projektionszentum aus der SVD Die Singulärwertzerlegung einer Projektionsmatrix P liefert drei Ergebnissmatrizen U, d und V . Wenn das untere, rechte Element von d Null ist, befindet sich in der rechten Spalte von V der gesuchte Nullvektor C.

Projektionszentrum aus den Unterdeteminanten von P In [22] ist gezeigt, dass sich C über die 3×3-Unterdeterminanten von P berechnen lässt, wobei $reduce\,(P, x)$ die 3×3-Untermatrix von P durch Weglassen der Spalte x angibt:

$$C = \begin{bmatrix} \det\left(reduce\,(P,1)\right) \\ -\det\left(reduce\,(P,2)\right) \\ \det\left(reduce\,(P,3)\right) \\ -\det\left(reduce\,(P,4)\right) \end{bmatrix} \tag{85}$$

Das Projektionszentrum muss danach auf $C^w = 1$ normiert werden.

Projektionszentrum aus Teilinvertierung von P Aus Gleichung 33 kann eine weitere Methode zur Projektionszentrumsbestimmung abgeleitet werden:

$$\begin{aligned} P &= [M\,|v] = KR\,[I\,|-C]] \\ &= [KR\,|-KRC]] \end{aligned}$$

$$\begin{aligned} M &= KR \\ v &= -MC \end{aligned} \tag{86}$$

$$\Longleftrightarrow C = -M^{-1}v$$

Hierbei ist zu beachten, dass M nicht invertierbar ist, wenn die Kamera im Unendlichen liegt. Für diesen Fall kann das Projektionszentrum über den Nullvektor von M bestimmt werden:

$$\begin{aligned} (U, d, V) &= svd\,(M) \\ C &= \left[V^{13}, V^{23}, V^{33}, 0\right]^T \end{aligned} \tag{87}$$

Auswahl des Verfahrens zur Bestimmung des Projektionszentrums Die drei Verfahren zeigen auf synthetischen Daten mit unterschiedlichem Rauschanteil keinen signifikanten Unterschied in der Qualität der Projektionszentrumsbestimmung. Da die Invertierung einer 3×3-Matrix direkt erfolgen kann, ist dieses Verfahren das schnellste, da es nur die Invertierung von M und eine Matrixmultiplikation benötigt. Des weiteren ist es von Vorteil, dass Kameras im Unendlichen gesondert betrachtet werden, da diese Projektionsmatrizen andere mathematische Eigenschaften besitzen. Daher wird im Folgenden das Kamerazentrum durch Teilinvertierung von P bestimmt.

11.5 Bündelblockausgleich

Das System soll photogrammetrischen Ansprüchen genügen, daher wird für die endgültige Kameraposition eine Ausgleichsrechnung mittels Bündelblockausgleich (BBA) durchgeführt. Das Gesamtsystem umfasst in der Regel mehrere Tausend Kameras mit Zehntausenden von 3D-Punkten und Hunderttausenden von Rückprojektionen. Dies stößt zum gegenwärtigen Zeitpunkt an die Grenzen der BBA und es müssen sogfältige Überlegungen für die sinnvolle Reduktion des Rechenaufwandes angesetzt werden.

Ein erster Ansatz zu Aufwandsminimierung besteht darin, die Lückenhaftigkeit (engl. *sparseness*) des Einflusses der zu optimierenden Parameter auszunutzen. Haben zwei Kameras keine gemeinsamen Punkte, beeinflussen sie sich auch nicht direkt, was in den benötigten Jakobimatrizen zu großen Gebieten führt, die Null sind. Ein frei verfügbares Framework, das diesen Umstand ausnutzt, ist das *Generic Sparse Bundle Adjustment* aus der Arbeit von [38] (Weblink im Anhang B). Allerdings steigt der Rechenaufwand bei steigender Kameraanzahl mit $O\,(n^2)$, was an der 2D-Struktur der Jakobimatrix liegt (vgl. Abbildung 16).

Als zweiter Ansatz kann die Zahl der Kameraparameter reduziert werden. Da die intrinsische Kalibrierung für jede Kamera während einer Sequenz als fix angenommen wird, kann der BBA mit fixer intrinsicher Kalibrierung rechnen, was die elf Parameter pro Kamera um fünf zu bestimmende Parameter reduziert. Parallel dazu müssen die intrinsischen Parameter von den jeweiligen Bildpunkten abgezogen werden (vgl. Gleichung 44). Dadurch haben die drei unterschiedlichen Kameras aus Sicht des BBA dieselbe intrinsiche Kalibrierung und können gleich behandelt werden.

Diese algorithmische Optimierung reicht, um ca. 2000 Kameras gleichzeitig zu bündeln. Allerdings dauert die Ausgleichsrechnung wegen der mangelnden Multiprozessorunterstützung einige Stunden. Sind zu viele Kameras in der Szene oder soll das Ergebnis schneller vorliegen, kann der Pfad in mehrere Blöcke unterteilt werden. Der entscheidende Parameter für die Qualität ist die Größe des Blocks. Sie sollte möglichst groß gewählt werden, muss aber auch an den zur Verfügung stehenden Speicher und die maximal tolerierbare Rechenzeit angepasst werden. In Abbildung 16 sind typische Zeiten eines BBA für unterschiedliche

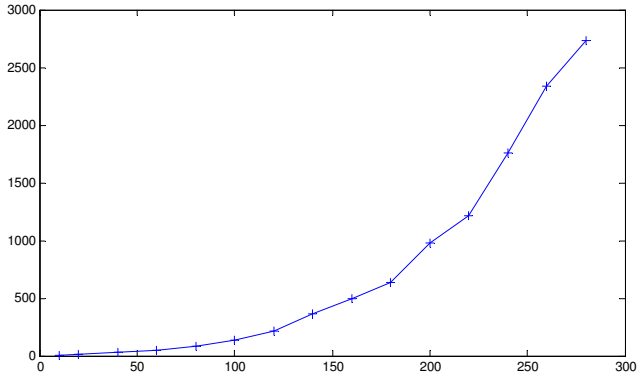


Abbildung 16: Laufzeiten für die Bündelausgleichsrechnungen

viele Kameratripel gezeigt. Während ein Ausgleich von nur 10 Kameratripeln durchschnittlich 3.3 Sekunden benötigt, werden für einen Ausgleich von 280 Kameratripeln schon 2735 Sekunden oder gut 45 Minuten benötigt. Für dieselbe Anzahl von Kameras braucht man in 10er-Schritten nur 94 Sekunden, hat aber die Zusammenhänge innerhalb des Netzwerkes sehr reduziert.

Hat man sich für eine Blockgröße entschieden, die die Gesamtanzahl der Kameras unterschreitet, müssen die einzelnen Blöcke wieder aneinandergefügt werden. Dies wird über eine gemeinsame Kamera realisiert. Die letzte Kamera des letzten Blocks ist identisch mit der ersten Kamera des zweiten Blocks, die als fix deklariert wird. Ein Problem dieser Technik liegt jedoch in der Schwierigkeit, den globalen Skalierungsfaktor gleich zu halten. In jedem BBA-Segment ist dieser globale Skalierungsfaktor unbestimmt und die nummerischen Verfahren halten diese Skalierung nicht zwingend konstant. Da nach einem BBA der Abstand der ersten und zweiten Kamera für jedes Triplet nicht mehr unbedingt gleich ist, ist es auch nicht ratsam, diesen Abstand wieder auf den vorgegebenen Wert zu setzen. Die Punktwolken aus den Einzelansichten passen danach nicht immer zueinander. Als gute Berechnungstechnik für die Skala hat sich eine Anpassung der 3D-Wolken aus den getrackten Punktkorrespondenzen erwiesen. Dabei wird angenommen, dass sich die Raumpunkte der letzten Kamera P^n_{last} des letzten Blocks von denen der zweiten Kamera P^2_{act} des aktuellen Blocks nur über einem Skalierungsfaktor unterscheiden:

$$|X_{last}| = \mu\,|X_{act}| \tag{88}$$

Dieser Skalierungsfaktor wird nun für alle Raumpunkte, die nahe an den Kameras liegen, mittels SVD bestimmt. Um ihn auch auf die Kameras anwenden zu können, muss die Referenzkamera P^1_{act} des aktuellen Blocks zunächst in den Ursprung transferiert werden. Die Skalierung lässt sich dann für alle Kameras berechnen:

$$P^i_{scaled} = P^i_{act} \begin{bmatrix} & P^1_{act} & \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 & \mu \\ 0 & 1 & 0 & \mu \\ 0 & 0 & 1 & \mu \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} & P^1_{act} & \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{89}$$

11.5.1 Bedingungen durch den Rahmen

Wie schon in Abschnitt 11.3.1 gezeigt, entstehen auch für die Ausgleichsrechnung durch den Rahmen Bedingungen. Zunächst scheinen die Bedingungen recht einfach: Durch einen starren Rahmen und eine bekannte Orientierung der Kameras kann an jeder zeitlichen Position der Referenzkamera die Position der anderen Kameras bestimmt werden. Allerdings würden diese Bedingungen voraussetzten, dass der Rahmen absolut starr wäre und sich niemals verziehen würde, und dass sich die Linsen nicht im Gehäuse bewegen würden. Dies ist jedoch

in der Realität nicht der Fall, da der Rahmen durch Verwindungen und Erschütterungen im Millimeterbereich variiert. Selbst die Länge der Basis auf dem Rahmen ist in einem gewissen Maß unbestimmt.

Wegen der großen Komplexität wurde auf die explizite Modellierung dieser Bedingungen für die Ausgleichsrechnung mit Varianzen und Kovarianzen, eine Linearisierung dieser Bedingungen sowie eine direkte Berechnung der Jakobimatrix verzichtet. Die Bedingungen werden nur zu Beginn des Bündelausgleichs eingeführt und der Ausgleichsrechnung wird gestattet, sie im Sinne der Fehlerminimierung zu verändern. Dazu werden zunächst die Kameraorientierungen auf dem Rahmen per Ausgleichsrechnung für alle Bilder bestimmt. Ausgehend von einer Orientierungsschätzung mit dem Verfahren aus Abschnitt 11.2 werden die Punktkorrespondenzen für jedes Bildtriplet unabhängig von den anderen Triplets trianguliert. Diese Raumpunkte stammen zwar von unterschiedlichen Positionen des Rahmens, allerdings ist die Orientierung der Kameras für alle Triangulationen gleich. Auch wenn derselbe Punkt in verschiedenen Tripeln trianguliert wurde, werden alle Triangulationen als unabhängige Messungen interpretiert und in die Ausgleichsrechnung integriert. Das Ergebnis beschreibt eine Ausgleichung der minimalen Bewegungen des Rahmens für die gesamte Szene und dient als Eingangsbedingung für die nachfolgende Ausgleichsrechnung des Gesamtpfades.

Zu Beginn wird eine initiale Pfadschätzung für die Referenzkamera mit der in Kapitel 11 beschriebenen Technik bestimmt. Im Anschluss daran werden für jede Position auf diesem Pfad die entsprechenden Positionen der zwei anderen Kameras durch Transformation nach Gleichung 7 bestimmt. Sodann werden für jedes verfolgte Punktmerkmal die Kameras zusammengetragen, in denen das Merkmal gefunden wurde. Sämtliche Korrespondenzen dieses Merkmals werden nun unter Verwendung von Gleichung 121 und den geschätzten Kameraorientierungen trianguliert. Diese Daten bilden den Startwert für den Bündelausgleich.

11.6 Ergebnisse

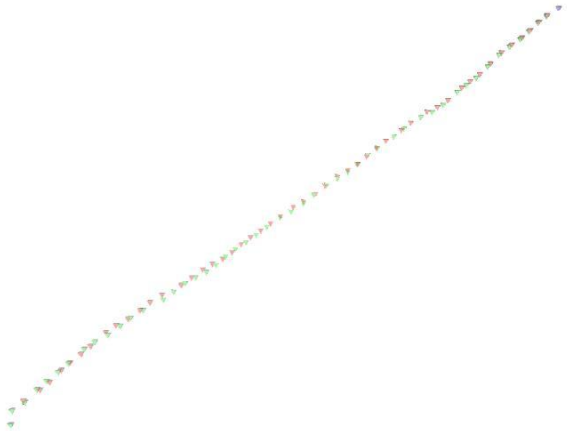
Die Qualität der Pfadschätzung ist von außerordentlicher Wichtigkeit für die Gesamtqualität des Systems. Zunächst wird das vorgestellte Verfahren für die Schätzung trifokaler Kamerapfade qualitativ mit bifokalen und monokularen Kamerapfadschätzungen verglichen, um zu zeigen, dass das bei monokularen Kamerapfadschätzungen auftretende Skalierungsproblem effektiv gelöst wurde. Sodann wird die trifokale Pfadrekonstruktion mit realen Messdaten quantitativ untersucht, um die absolute Genauigkeit des Systems bestimmen zu können.

11.6.1 Vergleich von monokularen, bifokalen und trifokalen Kamerapfaden

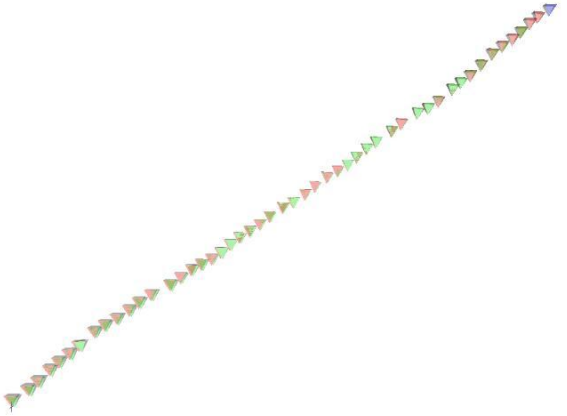
Für diesen Vergleich wurden die mit dem Trifokalfilter verfolgten Merkmale einer 200 Bilder langen Sequenz ausgewählt. Da die Skalierung bei der monokularen Pfadschätzung nur relativ zur ersten Kamerabasis erfolgen kann und daher der Vergleich mit realen Messwerten sehr aufwändig ist, wurde auf ein qualitatives Bewertungsverfahren zurückgegriffen. Die Bildreihenfolge der ausgewählten Szene wird nach 200 Bildern umgedreht, so dass sich die Kamera nach weiteren 200 Bildern wieder an der Startposition befindet. Je näher die Positionen des Hinwegs und Rückwegs beieinander liegen, desto genauer ist die Schätzung.

	Abstand in cm			Winkel in °		
	Ø	max.	σ	Ø	max.	σ
Monokular	74.28	148.38	41.48	0.0090	0.0393	0.0089
Bifokal	3.27	5.39	1.53	0.0142	0.0324	0.0067
Trifokal	2.71	6.36	1.79	0.0039	0.0109	0.0027

Tabelle 5: Vergleich von Pfadschätzungen mit unterschiedlicher Kameraanzahl



Monokularer Pfad



Bifokaler Pfad

Trifokaler Pfad

Abbildung 17: Kamerapfade

Damit die Daten für Hin- und Rückweg unabhängig voneinander sind, wurden auf dem Hinweg nur die graden Merkmalsindizes und auf dem Rückweg nur die ungeraden Indizes verwendet. Die Pfadschätzung mit nur einer Kamera benötigt eine Referenzskala für die erste Translation. Diese Referenz wurde aus den Ergebnissen der trifokalen Pfadschätzung übernommen, wodurch die Ergebnisse aller drei Pfadschätzungen direkt miteinander vergleichbar sind. Vor dem Vergleich der Kameraabstände wurde die initiale Pfadschätzung mit einem Bündelblockausgleich optimiert. Gemessen wurden der durchschnittliche und der maximale Abstand der theoretisch identischen Kameras in Zentimetern sowie die durchschnittliche und die maximale Abweichung der Ausrichtungswinkel in Grad. Zusätzlich wurde die Standard-

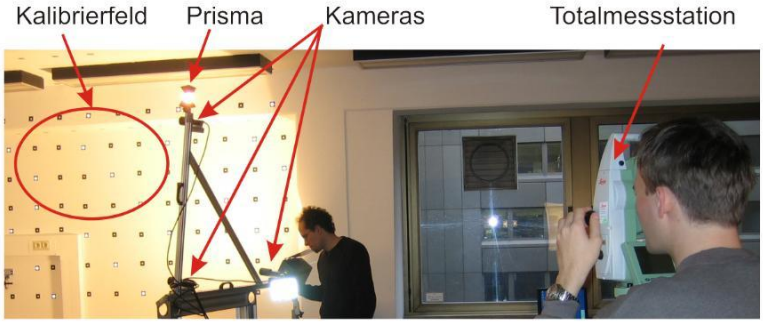


Abbildung 18: Rahmenkonfiguration für die Pfadmessung

abweichung dieser zwei Fehlermaße berechnet. Die Ergebnisse sind in Tabelle 5 zu sehen. Abbildung 17 zeigt die geschätzten Kamerapositionen in Form von Pyramiden für die drei Fälle. Die Startposition ist bei allen Pfaden unten links, der Wendepunkt befindet sich oben rechts. Die Positionen auf dem Hinweg sind mit roten Kameras und die auf dem Rückweg mit grünen Kameras eingezeichnet. Die Kameraposition am Wendepunkt ist blau dargestelt.

Die Ergebnisse aus Tabelle 5 bestätigen die Annahme, dass die Bedingungen durch einen Rahmen die Genauigkeit der Pfadschätzung signifikant verbessern. Die Abstände der bifokalen Schätzung verbessern sich durchschnittlich um Faktor 22, die der trifokalen Schätzung um Faktor 27. Auch die Standardabweichung ist bei den Mehrkamerasystemen deutlich geringer. Die geringfügig höhere Standardabweichung beim trifokalen Pfad liegt mit 2.4 Millimetern im vernachlässigbaren Bereich. Auch der höhere maximale Fehler von weniger als ein Zentimeter ist nicht gravierend, weil der durchschnittliche Fehler um 5,7mm geringer ist. Wie in [54] bereits beschrieben, kann die Kamerarotation viel genauer geschätzt werden als die Translation. Hier ist der monokulare Pfad im Durchschnitt sogar geringfügig genauer als der bifokale Pfad, jedoch aber nicht so exakt wie der trifokale Pfad.

Bei der visuellen Analyse von Abbildung 17 wird zunächst der monokulare mit dem bifokalen Pfad verglichen. Deutlich zu erkennen ist, dass die Kamerapositionen beim monokularen Pfad auf dem Hin- und Rückweg stark voneinander abweichen, während die Positionen beim bifokalen Pfad fast identisch sind und die Abweichung der roten und grünen Kameras nur bei den Positionen unten links zu erkennen ist.

Die Unterschiede zwischem bifokalen und trifokalem Pfad sind visuell nicht erkennbar, allerdings fällt bei beiden Pfaden auf, dass der Fehler in der Nähe der Startposition zunimmt. Hierbei sei angemerkt, dass der Bündelblockausgleich durch die Datensatzteilung nur die Hälfte der Punkte für jede Position verarbeiten konnte. Die Genauigkeit des vollständigen Datensatzes ist daher höher, was im folgenden Abschnitt gezeigt wird. Allerdings wird deutlich, dass die Qualität der Pfadrekonstruktion in hohem Maße von der Anzahl der verfolgten Punktmerkmale abhängt.

11.6.2 Evaluation mit Referentzdaten einer Totalmessstation

Um reale Daten zu verweden, mussten die Kamerapositionen für einen gesamten Pfad gemessen werden, um anschließend mit den geschätzten Positionen verglichen zu werden. Als Messinstrument für die Kamerabewegung wurde eine Totalmessstation (TMS) der Marke Leica mit Trackingmöglichkeit verwendet, die eine Messgenauigkeit von 3mm bei 7Hz Trackingfrequenz angegeben hat. Dazu wurde auf dem Rahmen ein Trackingprisma angebracht und mit Hilfe eines Kalibrierfeldes eingemessen (vgl. Abbildung 18).

Die TMS kann keine Rotation des Prismas detektieren, sondern nur dessen Position. Um das Prisma gut sichtbar zu halten, wurde es oberhalb der dritten Kamera angebracht, die bei der Pfadbewertung als Referenz verwendet wurde. Durch den Abstand zwischen Prisma und Kamera 3 von 19.36cm kommt es bei Bewegung des Kamerarahmens zu einer

Taumelbewegung zwischen Prisma und Kamera 3. Eine Rotation um das Projektionszentrum der Kamera 3 resultiert in einer nicht messbaren Taumelbewegung (vgl. Abbildung 19), deren Einflussgröße in Tabelle 6 aufgeführt ist. Nimmt man bei der Kameraführung eine Taumelbewegung von 2° an, können die Kamerapositionen aus den Messwerten der TMS nicht genauer als 0.67cm plus 0.3cm Abweichung als Ergebnis der Messungenauigkeit, also ca. 1cm, bestimmt werden. Die Länge der Kamerabasis zwischen Kamera 1 und Kamera 2 wurde über das Kalibrierfeld bestimmt und beträgt 890.8mm. Des weiteren wurden die intrinsische Kalibrierung und die Verzeichnungsparameter der Kameras über das Kalibrierfeld ermittelt.

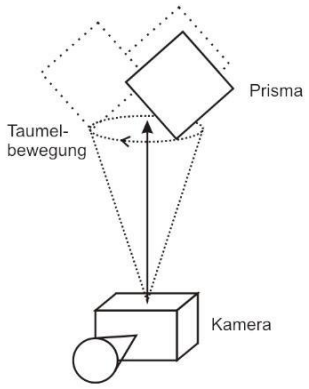


Abbildung 19: Taumeln des Prismas bei Rotation um die y-Achse

	x-Achse	y-Achse	z-Achse
2°	0.66cm	0.67cm	0.16cm
5°	1.64cm	1.67cm	0.40cm
10°	3.28cm	3.67cm	0.80cm

Tabelle 6: Taumelfehler des Prismas

Als Objekt wurde die Fassade des Ernst-Reuter-Hauses in Berlin gewählt. Die Fassade hat eine Länge von 111m, eine Traufhöhe von 14.5m beziehungsweise 18m in der Gebäudemitte und eine Tiefe von 5.1m zwischen den einzelnen Fassadenteilen. Der Aufnahmeabstand betrug 25m. Die Aufnahme umfasst 2100 Bildtripel mit einer Aufnahmefrequenz von 14Hz. Der rekonstruierte Pfad und die für das Tracking verwendeten Punkte sind in Abbildung 20 gezeigt.

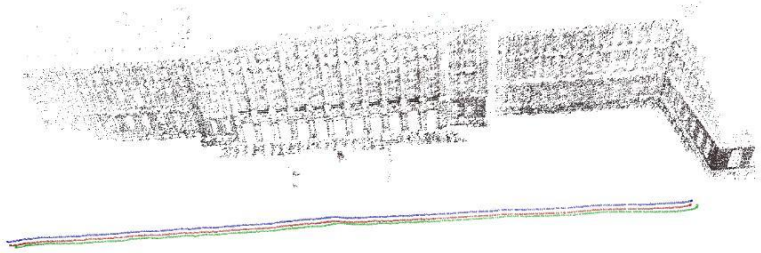
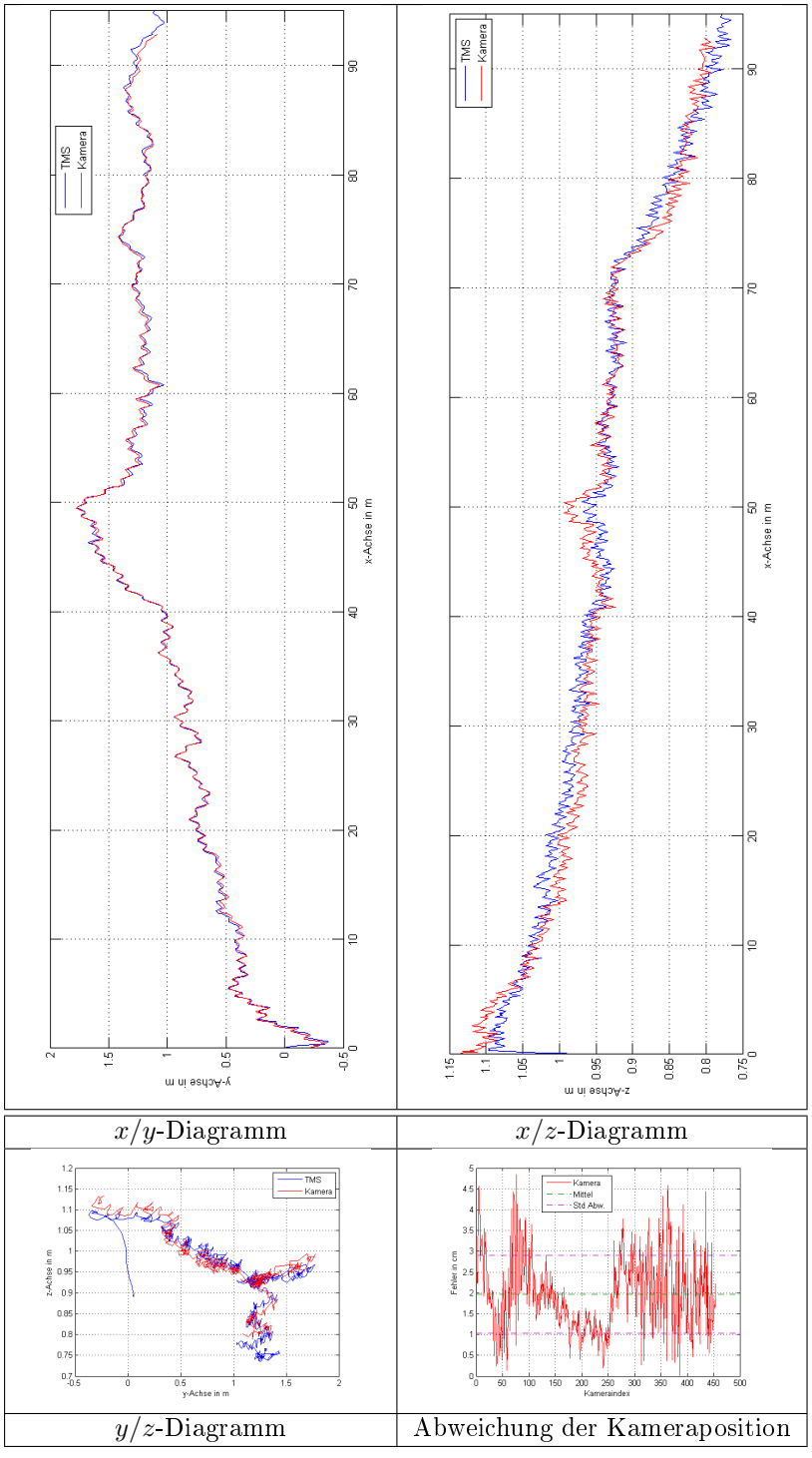


Abbildung 20: Kamerapfad entlang des Ernst-Reuter-Hauses

Um die Messung der Prismabewegung mit dem geschätzten Kamerapfad vergleichen zu können, müssen beide Pfade in dasselbe Koordinatensystem übertragen werden. Da nur die räumliche Beziehung zwischen den Kameras und dem Prisma bekannt ist, muss die Drehbewegung der zwei Bewegungspfade noch angeglichen werden. Nach einer groben manuellen Ausrichtung der beiden Pfade erfolgte die endgültige Anpassung der Pfade aneinander über einen *Iterative Closest Point* (ICP)-Algorithmus [57]. Der verwendete Code ist als Weblink

in Anhang B referenziert. Um den BBA über den gesamten Pfad ausführen zu können, wurde nur jedes vierte Kameratriplet für den Bündelausgleich verwendet. Es wurden 526 Kamerapositionen und 1008 Prismapositionen verglichen. Für die Zuordnung der Prismapositionen zu den Kamerapositionen wurden die Prismapositionen linear interpoliert (vgl. Abbildung 22). Als Fehlermaß wurde für die Auswertung und für das Optimierungskriterium des ICP der minimale Abstand eines Messpunktes (rot) zur Verbindungslinie der zwei nächstgelegenen Prismapositionen (blau) verwendet.



x/y-Diagramm | x/z-Diagramm

y/z-Diagramm | Abweichung der Kameraposition

Abbildung 21: Vergleich des Kamerapfads mit den Trackingdaten

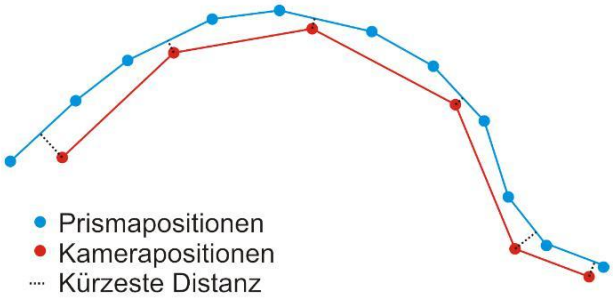


Abbildung 22: Berechnung des Fehlermaßes für die Kameraposition

Pfaddimension		Fehler	
x-Achse	92.684m	Min	0.03cm
y-Achse	2.121m	Max	4.85cm
z-Achse	0.346m	Mittel	1.97cm
Länge	92.694m	σ	0.94cm

Tabelle 7: Auswertung des Kamerapfades

Die Vergleichsergebnisse des aufgenommenen Pfades sind in Tabelle 7 und Abbildung 21 zu sehen. Der rote Kamerapfad weicht auf einer Länge von 92.694m maximal um 4.85cm ab, was einem Verhältnis von maximal 0.0523% entspricht. Der durchschnittliche Fehler liegt unter 2cm und die Standardabweichung ist mit 0.94cm im Bereich der globalen Messgenauigkeit.

Die Genauigkeit der Pfadschätzung ist über die gesamte Pfadlänge sehr gleichmäßig (vgl. Abbildung 21 rechts unten). Die maximale Abweichung vom Mittelwert beträgt nur 2.88cm und entspricht damit etwas mehr als der dreifachen Standardabweichung. Am Ende der ersten Hälfte des Kamerapfades sinkt der Fehler auffällig, während er am Anfang und zum Ende hin vergleichsweise stark um den Mittelwert pendelt. Dies ist damit zu erklären, dass in dem Bereich der Szene mit dem niedrigen Fehler die Kamerageschwindigkeit etwas verringert wurde und der Abstand zur TMS in diesem Bereich am geringsten war. Zusätzlich befand sich dieser Fassadenbereich ca. 5m näher am Aufnahmesystem als die Fassadenbereiche mit höherem Fehler.

12 Rektifizierung

Die Korrespondenzsuche für die 3D-Rekonstruktion erfolgt entlang der Epipolarlinien. Diese Linien verlaufen in der Regel nicht parallel zum Bildraster, weswegen jeder Kandidat auf den Epipolarlinien durch seine benachbarten Bildpunkte interpoliert werden müsste. Dies müsste für jede Position im Sekundärbild mehrfach durchgeführt werden, da mehrere Epipolarlinien diese Bildposition in leicht unterschiedlichen Subpixelpositionen schneiden. Diese Berechnung ist aufwändig und kann auch beim symmetrischen Stereo zu unerwünschten Rastereffekten führen.

Wie in Kapitel 8 erwähnt, wird daher die Korrespondenzsuche auf Stereonormalbildern durchgeführt. Werden Korrespondenzen in drei Bildern, deren Projektionszentren nicht auf einer Linie liegen, gleichzeitig gesucht, kann das Konzept der Stereonormalbilder auf den Dreibildfall erweitert werden, bei dem die Epipolarlinien des horizontalen Paares parallel zu den Bildzeilen und die Epipolarlinien des vertikalen Bildpaares parallel zu den Bildspalten verlaufen. Die Transformation eines Bildtripels in diese Konfiguration heißt trinokulare Rektifizierung. Die vorgestellte Technik basiert auf der Arbeit von [67] und stellt eine Verfeinerung der vorausgegangenen Arbeit [25] insbesondere bei der Sortierung der Kameras dar. Dazu wird für die Bilder eine virtuelle Projektionsebene berechnet, die parallel zu der durch die drei Projektionszentren bestimmten Kameraebene liegt (vgl. Abbildung 23). Es ist sofort ersichtlich, dass die drei Projektionszentren dabei nicht auf einer Geraden liegen

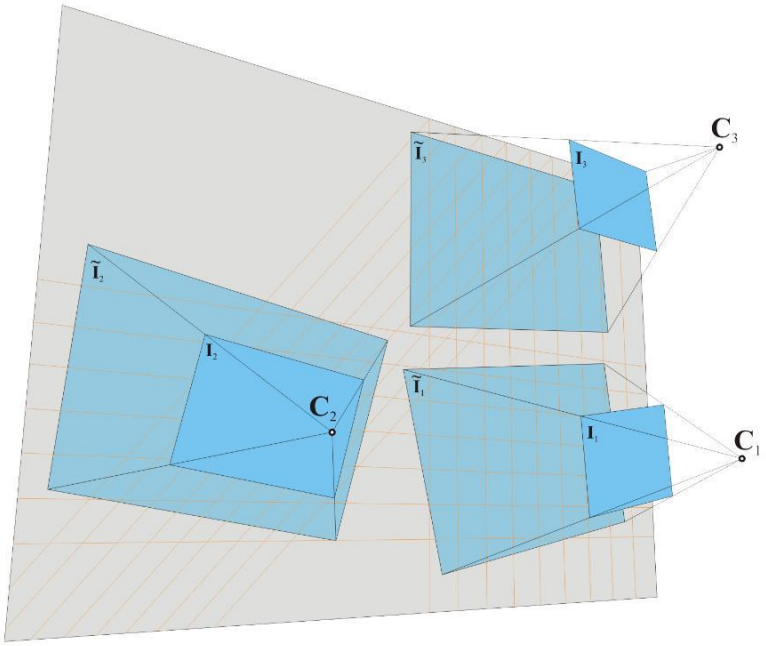


Abbildung 23: Trinokulare Rektifizierung

dürfen, weil sonst die Kameraebene nicht bestimmt werden kann. Ferner ist dieses Verfahren im Allgemeinen nicht für die Rektifizierung von mehr als drei Bildern geeignet, weil das vierte Projektionszentrum nicht mehr auf einer Ebene mit den anderen liegen muss.

Im Folgenden werden die Voraussetzungen für die trinokularen Rektifizierung sowie ihre Eigenschaften und lineare Berechnung beschrieben. Dieses Verfahren arbeitet ausschließlich mit Bildkorrespondenzen, es werden Strategien zur Bestimmung freier Parameter vorgestellt und die Qualität der Rektifizierung wird experimentell bestätigt.

12.1 Trifokalmodell

Für eine trinokolare Rektifizierung müssen die drei Kameras nicht kalibriert werden. Es reicht aus, wenn das Bildtripel geradentreu ist (vgl. Abschnitt 9.3) und sechs korrespondierende Punkte in allen drei Bildern mittels der Technik aus Kapitel 10 bestimmt sind, die nicht auf einer Ebene liegen, um den Trifolaltensor berechnen zu können, wie es in |22| beschrieben ist. Das Vorhandensein von Bildkorrespondenzen impliziert auch, dass die drei Bilder einen gemeinsamen Überlappungsbereich haben. Wie bereits angemerkt dürfen die Projektionszentren nicht kollinear sein und die Projektionszentren sollten nicht in den jeweils anderen zwei Bildern zu sehen sein. Ansonsten würde, wie im Zweibildfall, eine lineare Rektifizierung die Bilder ins Unendliche verzerren.

Da Kamerakalibrierung und -orientierung nicht benötigt werden, ist es möglich, die Aufnahmegeometrie und die Linseneinstellungen wie Blende und Fokus optimal auf das aufzunehmende Objekt einzustellen, die Aufnahmen zu tätigen und erst danach die Rektifizierung zu bestimmen. Des weiteren kann die Aufnahmegeometrie zum Transport verändert werden, ohne dass die Qualität der Rektifizierung beeinflusst wird. Die Pixelgröße der verwendeten Kameras beträgt nur 3,5µm, daher können schon leichte Stöße, die selbst bei vorsichtigem Transport fast unvermeidbar sind, oder Temperaturschwankungen die Aufnahmegeometrie um mehrere Pixel verändern. In diesem Kapitel sind die Variablen, Vektoren und Matrizen der rektifizierten Bilder mit Tilden (˜) gekennzeichnet.

12.1.1 Modellhafte Annahme der Kamerapositionierung

Die Erweiterung des Stereonormalfalls auf Bildtripel sieht vor, dass die Bilder in einem rechtwinkligen, gleichseitigen Dreieck positioniert werden, das der Form des gekippten Buch-

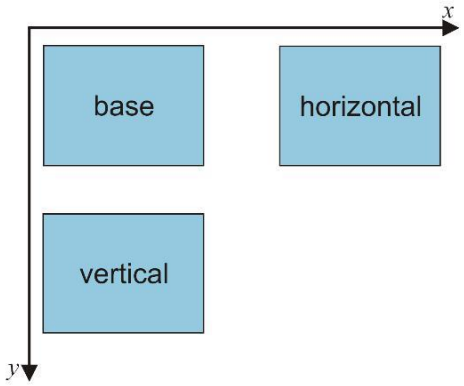


Abbildung 24: Ideale Rektifizierung

stabens L ähnelt (vgl. Abbildung 24). Diese Anordnung erhält den Stereonormalbildfall für das horizontale Bildpaar, das heißt, die Bildzeilen korrespondieren immer noch mit den Epipolarlinien. Für das vertikale Bildpaar gilt, dass die Bildspalten mit den Epipolarlinien übereinstimmen. Dafür müssen die Epipole aller Bilder im Unendlichen liegen und alle Epipolargeraden zwischen dem Sekundär- und dem Tertiärbild parallel zueinander liegen. Da Verschiebungen und Scheerungen des horizontalen Bildes entlang der horizontalen Achse sowie des vertikalen Bildes entlang der vertikalen Achse diese Eigenschaften nicht verändern, kann sichergestellt werden, dass der Anstieg der Epipolargeraden zwischen diesem Bildpaar -1 beträgt. Daher gilt für alle Korrespondenzen, dass die Disparitäten zwischen horizontalem und vertikalem Bild angeglichen werden können:

$$\tilde{I}_1(x,y) = \tilde{I}_2\left(x + \tilde{D}(x,y), y\right) = \tilde{I}_3\left(x, y + \tilde{D}(x,y)\right) \tag{90}$$

Bildtripel mit dieser Eigenschaft haben besondere, nämlich im Unendlichen liegende Epipole:

$$\begin{aligned} \tilde{e}_{bh} = \tilde{e}_{hb} &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T \\ \tilde{e}_{bv} = \tilde{e}_{vb} &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T \\ \tilde{e}_{hv} = \tilde{e}_{vh} &= \begin{bmatrix} -1 & 1 & 0 \end{bmatrix}^T \end{aligned} \tag{91}$$

Die Fundamentalmatrizen dieser Bilder lassen sich durch die Epipole bestimmen und haben nur wenige von Null abweichende Elemente:

$$\begin{aligned} \tilde{F}_{bh} &= [\tilde{e}_{bh}]_\times = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \\ \tilde{F}_{bv} &= [\tilde{e}_{bv}]_\times = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \\ \tilde{F}_{hv} &= [\tilde{e}_{hv}]_\times = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix} \end{aligned} \tag{92}$$

12.1.2 Rektifizierung

Ziel der Rektifizierung ist es, drei 3×3-Homographien H_b, H_h und H_v zu finden, welche das reale Bildtripel so transformieren, dass die Fundamentalmatrizen des Tripels Gleichung 92 entsprechen. Die Fundamentalmatrizen des nicht rektifizierten Bildtripels können aus dem Trifokaltensor extrahiert werden, der durch die sechs Bildkorrespondenzen bestimmt ist [22].

Diese drei Fundamentalmatrizen sind nicht unabhängig voneinander und verfügen insgesamt nur über 18 Freiheitsgrade, weil zusätzlich Gleichung 43 gelten muss.

Die Rektifizierungsmatrizen transformieren jeden Bildpunkt x_i auf einen neuen Bildpunkt $\tilde{x}_i$ für alle $i\epsilon\{b,h,v\}$:

$$H_i = \begin{bmatrix} u_i^T \\ v_i^T \\ w_i^T \end{bmatrix} = \begin{bmatrix} u_{i1} & u_{i2} & u_{i3} \\ v_{i1} & v_{i2} & v_{i3} \\ w_{i1} & w_{i2} & w_{i3} \end{bmatrix} \tag{93}$$
$$\tilde{x}_i = H_i x_i$$

Für alle drei Bildpaare muss sowohl im rektifizierten als auch im nicht rektifizierten Fall Gleichung 39.3 gelten:

$$\begin{aligned} x_h^T F_{bh} x_b = \tilde{x}_h^T \tilde{F}_{bh} \tilde{x}_b = 0 \\ x_v^T F_{bv} x_b = \tilde{x}_v^T \tilde{F}_{bv} \tilde{x}_b = 0 \\ x_v^T F_{hv} x_v = \tilde{x}_v^T \tilde{F}_{hv} \tilde{x}_h = 0 \end{aligned} \tag{94}$$

Setzt man Gleichung 93 in Gleichung 94 ein und berücksichtigt, dass Fundamentalmatrizen bis auf drei Skalierungsfaktoren λ_j bestimmt sind, erhält man:

$$\begin{aligned} H_h^T \tilde{F}_{bh} H_b = \lambda_1 F_{bh} \\ H_v^T \tilde{F}_{bv} H_b = \lambda_2 F_{bv} \\ H_v^T \tilde{F}_{hv} H_v = \lambda_3 F_{hv} \end{aligned} \tag{95}$$

Da die Fundamentalmatrizen der rektifizierten Bilder sehr einfach sind, kann man durch Einsetzen der Gleichungen 92 und 93 in Gleichung 95 eine vereinfachte Form erhalten:

$$\begin{aligned} w_h v_b^T - v_h w_b^T = \lambda_1 F_{bh} \\ u_v w_b^T - w_v u_b^T = \lambda_2 F_{bv} \\ (u_v + v_v)\, w_h^T - w_v\, (u_h + v_h)^T = \lambda_3 F_{hv} \end{aligned} \tag{96}$$

Jede Fundamentalmatrix hat nur Rang 2, daher ist Gleichungssystem 96 unterbestimmt. Es müssen also zusätzliche Bedingungen gefunden werden. Ziel der Rektifizierung ist es, die sechs Epipole e_{bh}, e_{hb}, e_{vb}, e_{bv}, e_{hv} und e_{vh} mittels der Homographien ins Unendliche zu transformieren. Um das zu gewährleisten, müssen die Homographien die homogene Komponente beider Epipole der anderen beiden Bilder auf Null zu setzen. Für die w-Vektoren der Homographien gilt also:

$$\begin{aligned} w_b e_{bh} = w_b e_{bv} = 0 \\ w_h e_{hb} = w_h e_{hv} = 0 \\ w_v e_{vb} = w_v e_{vh} = 0 \end{aligned} \tag{97}$$

Mathematisch bedeutet dies, dass die w-Vektoren jeweils senkrecht zu zwei Epipolen stehen. Ein Vektor, der senkrecht zu zwei anderen Vektoren steht, kann durch das Kreuzprodukt dieser Vektoren bestimmt werden:

$$\begin{aligned} w_b = e_{bh} \times e_{bv} \\ w_h = e_{hb} \times e_{hv} \\ w_v = e_{vb} \times e_{vh} \end{aligned} \tag{98}$$

Die Epipole können mittels SVD aus den Fundamentalmatrizen extrahiert werden [22]. Die homogene Komponente der w-Vektoren muss auf Eins normiert werden. Somit hat Gleichung 96 nur noch 24 unbestimmte Parameter. Da die Fundamentalmatrizen 18 Freiheitsgrade bestimmen, bleibt ein Gleichungssystem mit sechs Freiheitsgraden bestehen. Daher werden die 27 linearen Gleichungen aus 96 explizit aufgeschrieben und sechs Variablen auf konstante Werte gesetzt. Analog zu [67] werden die Parameter λ_1, λ_2 und λ_3 auf Eins und die Parameter u_{b3}, v_{h3}und v_{v3} auf Null gesetzt.

Die Gleichungen sind somit linear lösbar und es können vorläufige Rektifizierungsmatrizen H_i^* berechnet werden:

$$
\begin{aligned}
H_b^* &= \begin{bmatrix} w_{b1}F_{bv}^{33} - F_{bv}^{31} & w_{b2}F_{bv}^{33} - F_{bv}^{32} & 0 \\ F_{bh}^{31} & F_{bh}^{32} & F_{bh}^{33} \\ w_{b1} & w_{b2} & 1 \end{bmatrix} \\
H_h^* &= \begin{bmatrix} w_{h1}\left(F_{bv}^{33} - F_{bh}^{33}\right) + F_{bh}^{13} - F_{hv}^{31} & w_{h2}\left(F_{bv}^{33} - F_{bh}^{33}\right) + F_{bh}^{23} - F_{hv}^{32} & F_{bv}^{33} - F_{hv}^{33} \\ w_{h1}F_{bh}^{33} - F_{bh}^{13} & w_{h2}F_{bh}^{33} - F_{bh}^{23} & 0 \\ w_{h1} & w_{h2} & 1 \end{bmatrix} \\
H_v^* &= \begin{bmatrix} F_{bv}^{13} & F_{bv}^{23} & F_{bv}^{33} \\ w_{v1}\left(F_{bv}^{33} - F_{hv}^{33}\right) + F_{hv}^{13} - F_{bv}^{13} & w_{v2}\left(F_{bv}^{33} - F_{hv}^{33}\right) + F_{hv}^{23} - F_{bv}^{23} & 0 \\ w_{v1} & w_{v2} & 1 \end{bmatrix}
\end{aligned}
\tag{99}
$$

Diese Homographien rektifizieren bereits die ursprünglichen Bilder, führen jedoch zu starken Verzerrungen und großen Translations- und Skalierungsunterschieden. Daher werden die vorher fixierten Parameter wieder in die Gleichung eingefügt. Um die Parameter besser interpretieren zu können, werden sie wie folgt umgewandelt:

$$
\begin{aligned}
s_1 = u_{b3} \qquad s_2 = v_{h3} \qquad s_3 = v_{v3} - v_{h3} \\
\alpha_1 = \lambda_1/\lambda_3 \qquad \alpha_2 = \lambda_2/\lambda_3 \qquad \alpha_3 = \lambda_3
\end{aligned}
\tag{100}
$$

Die Rektifizierung bestimmt sich also durch:

$$
\begin{aligned}
H_b &= \begin{bmatrix} 1 & 0 & s_1 \\ 0 & 1 & s_2 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \alpha_3 & 0 & 0 \\ 0 & \alpha_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \alpha_2 & 0 & 0 \\ 0 & \alpha_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot H_b^* \\
H_h &= \begin{bmatrix} 1 & 0 & s_1 + s_3 \\ 0 & 1 & s_s \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \alpha_3 & 0 & 0 \\ 0 & \alpha_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1-\alpha_1 & F_{bv}^{33}(\alpha_2 - 1) \\ 0 & \alpha_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot H_h^* \\
H_v &= \begin{bmatrix} 1 & 0 & s_1 \\ 0 & 1 & s_2 + s_3 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \alpha_3 & 0 & 0 \\ 0 & \alpha_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \alpha_2 & 0 & 0 \\ 1-\alpha_2 & 1 & F_{bv}^{33}(\alpha_2 - 1) \\ 0 & 0 & 1 \end{bmatrix} \cdot H_v^*
\end{aligned}
\tag{101}
$$

12.1.3 Freie Parameter

Für eine geeignete Rektifizierung müssen passende Zahlenwerte für die freien Parameter gefunden werden. Dazu werden zusätzliche Bedingungen an die rektifizierten Bilder gestellt, die theoretisch nicht zwingend erfüllt werden müssen, aber von praktischer Bedeutung sind:

- Die Bildgröße soll gleich bleiben
- Die Disparitäten der Bildkorrespondenzen sollen symmetrisch um Null liegen
- Die Trapezverzerrung der Bilder soll möglichst gering sein

Diese Bedingungen ergeben sich aus den Anforderungen, die bei der anschließenden Korrespondenzsuche anfallen. Der erste Punkt ist offensichtlich: Werden die Bilder zu stark verkleinert, gehen Informationen verloren. Werden die Bilder zu stark vergrößert, treten starke Interpolationsfehler auf. Da die Korrespondenzsuche in den Bildern symmetrisch durchgeführt wird, führen stark asymmetrische Suchbereiche zu unerwünschten Effekten im Randbereich, weil unterschiedliche Bereiche nicht von beiden Seiten berechnet werden können. Der letzte Punkt ist wichtig, wenn fensterbasierte Ähnlichkeitsfunktionen verwendet werden, insbesondere für die Subpixelinterpolation. Da eine dynamische Anpassung der Fensterform algorithmisch sehr rechenaufwändig ist, würden stark verzerrte Bilder zu fehlerhaft interpolierten Ergebnissen führen.

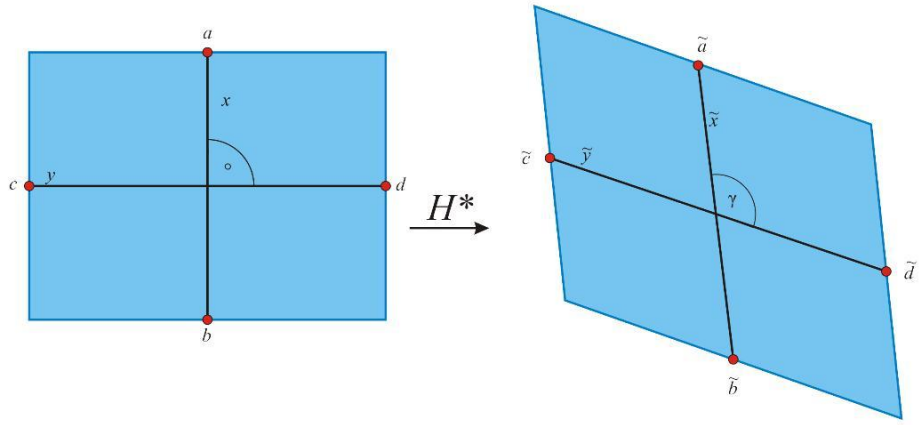


Abbildung 25: Scheerung der Rektifizierung

Die Parameter α_1 und α_2 sind Scheerungsfaktoren des horizontalen und vertikalen Bildes, die die Trapezverzerrung beeinflussen können. Für möglichst unverzerrte Bilder wird versucht, das durch die Bildhalbierenden gebildete Kreuz orthogonal zu halten. Daraus ergibt sich eine berechenbare Bedingung für die Scheerungsfaktoren. Es wird nur die Bestimmung von α_1 gezeigt, weil die Rechnung für α_2 analog verläuft.

Zuerst werden vier Punkte auf der Hälfte der Bildseiten a, b, c und d mit H_h^* transformiert und die Vektoren $\widetilde{x} = b^* - a^*$ und $\widetilde{y} = d^* - c^*$ berechnet (vgl. Abbildung 25).

$$\begin{array}{cccc} a = \begin{bmatrix} W/2 \\ 0 \\ 1 \end{bmatrix} & b = \begin{bmatrix} W/2 \\ H \\ 1 \end{bmatrix} & c = \begin{bmatrix} 0 \\ H/2 \\ 1 \end{bmatrix} & d = \begin{bmatrix} W \\ H/2 \\ 1 \end{bmatrix} \\ a^* = H_h^* & b^* = H_h^* b & c^* = H_h^* c & d^* = H_h^* d \end{array} \tag{102}$$

Ziel ist es nun, den Winkel γ durch Scheerung auf 90° zu transformieren. Dazu wird mittels des Scheerungsanteils S_h der Gleichung 101 versucht, das Skalarprodukt der Vektoren $\widetilde{x}$ und $\widetilde{y}$ auf Null zu setzen:

$$\begin{gathered} S_h = \begin{bmatrix} 1 & 1-\alpha_1 & 0 \\ 0 & \alpha_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ (S_h\widetilde{x})^T \cdot (S_h\widetilde{y}) = 0 \\ \alpha_1{}^2 + \alpha_1 \left(\frac{-\tilde{x}_x\tilde{y}_y - \tilde{x}_y\tilde{y}_x - 2\tilde{x}_y\tilde{y}_y}{2\tilde{x}_y\tilde{y}_y} \right) + \left(\frac{\tilde{x}_x\tilde{y}_x + \tilde{x}_x\tilde{y}_y + \tilde{x}_y\tilde{y}_x + \tilde{x}_y\tilde{y}_y}{2\tilde{x}_y\tilde{y}_y} \right) = 0 \end{gathered} \tag{103}$$

Hierbei ist zu beachten, dass der Wert $F_{bv}^{33}(\alpha_2 - 1)$ auf Null gesetzt werden kann, weil er ausschließlich eine Translation in x-Richtung verursacht und daher für die Scheerung irrelevant ist. Die quadratische Gleichung 103.3 hat meißt zwei Lösungen, wobei das Ergebniss mit dem kleineren Absolutwert $|\alpha_1|$ vorzuziehen ist.

Der Parameter α_3 ist ein globaler Skalierungsfaktor, der die Größe aller drei Bilder beeinflusst. Nachdem die Scheerungsparameter α_1 und α_2 bestimmt wurden, kann ein geeigneter Wert für diesen Parameter berechnet werden. Dazu wird versucht, die Länge der Bilddiagonale im Referenzbild konstant zu halten:

$$a = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T \qquad b = \begin{bmatrix} W & H & 1 \end{bmatrix}^T$$

$$\tilde{a} = \begin{bmatrix} \alpha_2 & 0 & 0 \\ 0 & \alpha_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot H_b^* \cdot a \quad \tilde{b} = \begin{bmatrix} \alpha_2 & 0 & 0 \\ 0 & \alpha_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot H_b^* \cdot b \tag{104}$$

$$\alpha_3 = \frac{|\tilde{a} - \tilde{b}|}{|a - b|}$$

Mit dem Parameter s_3 kann der Wertebereich der Disparitäten verschoben werden. Da die Bilder bei der Korrespondenzsuche symmetrisch bearbeitet werden, bietet es sich an, den Wertebereich so anzupassen, dass der Absolutwert der minimalen und maximalen Disparitäten gleich ist. Unter der Annahme, dass die für die Trifokaltensorberechnung verwendet Punktkorrespondenzen repräsentativ für das gesamte Bild sind, werden zunächst alle Punktkorrespondenzen x_i und x_i' mit $i \epsilon \{1, \cdots, N\}$ ohne Translationanpassung rektifiziert.

$$\begin{aligned} \widetilde{x}_i &= \begin{bmatrix} \alpha_3 & 0 & 0 \\ 0 & \alpha_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \alpha_2 & 0 & 0 \\ 0 & \alpha_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot H_b^* \cdot x_i \\ \widetilde{x}_i' &= \begin{bmatrix} \alpha_3 & 0 & 0 \\ 0 & \alpha_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1-\alpha_1 & F_{bv}^{33}(\alpha_2 - 1) \\ 0 & \alpha_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot H_h^* \cdot x_i' \end{aligned} \tag{105}$$

Da die Disparitäten im vertikalen Bildpaar gleich sind, müssen sie nicht berechnet werden. Nun werden minimale und maximale Disparität dieser Korrespondenzen bestimmt und der Parameter s_3 so gesetzt, dass die Mitte des Wertebereiches der Disparitäten bei Null liegt:

$$\begin{aligned} &minDisp = min\left(\widetilde{x}_i'^x - \widetilde{x}_i^x\right) \\ &maxDisp = max\left(\widetilde{x}_i'^x - \widetilde{x}_i^x\right) \\ &s_3 = -\left(\frac{maxDisp - minDisp}{2} + minDisp\right) = -\frac{maxDisp + minDisp}{2} \end{aligned} \tag{106}$$

Zuletzt werden die globalen Translationsparameter $s_1$und s_2 berechnet. Damit die Bildkoordinaten positiv sind, sollte der jeweils kleinste Wert in x- und y-Richtung auf Null gesetzt werden. Da die Bilder rechteckig sind und die Transformation linear ist, werden die Extremwerte durch die Bildecken begrenzt. Zunächst werden die Parameter $s_1$und s_2 auf Null gesetzt und dann alle vier Ecken der drei Bilder mit sämtlichen bereits bestimmten Parametern rektifiziert. Nun werden die kleinsten Werte x_{min} und y_{min} aus diesen zwölf transformierten Eckpunkten bestimmt. Um in der finalen Rektifizierung diese Werte auf Null zu setzen, werden die Translationsparameter auf $s_1 = -x_{min}$ und $s_2 = -y_{min}$ gesetzt.

12.1.4 Abschätzung des Suchbereiches

Die Korrespondenzen, die zur Bestimmung des Trifokaltensors verwendet wurden, können nun mit den vollständig bestimmten Homographien rektifiziert werden. Mit Hilfe dieser rektifizierten Korrespondenzen und Gleichung 106 können die maximalen und minimalen Disparitäten der Korrespondenzen bestimmt werden. Da die Disparitäten in den drei Bildern nach der Rektifizierung gleich sind, reicht es aus, diese Berechnung nur auf einem Bildpaar durchzuführen. Somit ist automatisch der Suchbereich für diese Korrespondenzen für das anschließende SGM bestimmt. Da allerdings sehr häufig auch Tiefenbereiche untersucht werden sollen, die etwas außerhalb dieses Suchbereiches liegen, wird eine Vergrößerung dieses Bereichs um 20% empfohlen:

$$\begin{aligned} &\widetilde{x}_i = H_b \cdot x_i \\ &\widetilde{x}_i' = H_h \cdot x_i' \\ &\widetilde{x}_i'' = H_v \cdot x_i'' \\ &minDispSGM = 1.2min\left(\widetilde{x}_i'^x - \widetilde{x}_i^x\right) = 1.2min\left(\widetilde{x}_i''^y - \widetilde{x}_i^y\right) \\ &maxDispSGM = 1.2max\left(\widetilde{x}_i'^x - \widetilde{x}_i^x\right) = 1.2max\left(\widetilde{x}_i''^y - \widetilde{x}_i^y\right) \end{aligned} \tag{107}$$

Ø−Scheerung der Ecken: 4.07° Ø−Scheerung der Ecken: 1.46°

Abbildung 26: Scheerung durch trifokale Mehrdeutigkeiten

12.1.5 Sortierung der Kameras

Bei manchen Aufnahmen kann nicht eindeutig bestimmt werden, welches Bild als Referenz dienen soll, beispielsweise wenn die Aufnahmepositionen ein gleichseitiges Dreieck bilden. Um vollautomatisch das beste Referenzbild auszuwählen, müssen die Kameras daher eventuell umsortiert werden.

Als Kriterium für eine gute Sortierung werden die Verzerrungen der Bildecken verwendet. Die Bildecken sollten nach einer Rektifizierung noch möglichst rechte Winkel haben. Die Abweichung err_α des Eckwinkels errechnet sich aus den drei rektifizierten Eckpunkten $\widetilde{x}_1$, $\widetilde{x}_2$ und $\widetilde{x}_3$:

$$\begin{aligned} h_1 &= \tfrac{\widetilde{x}_2 - \widetilde{x}_1}{|\widetilde{x}_2 - \widetilde{x}_1|} \\ h_2 &= \tfrac{\widetilde{x}_3 - \widetilde{x}_1}{|\widetilde{x}_3 - \widetilde{x}_1|} \\ err_\alpha &= \left|\tfrac{\pi}{2} - \left|\arccos\left(h_1^T h_2\right)\right|\right| \end{aligned} \tag{108}$$

Für die Sortierung werden alle sechs Permutationen der Bilder rektifiziert und die durchschnittliche Abweichung der vier rektifizierten Eckwinkel nach Gleichung 108 berechnet. Die Permutation mit der geringsten Abweichung sollte verwendet werden.

12.1.6 Rektifizierungstransformation

Die so berechneten Rektifizierungshomographien genügen den meisten Ansprüchen. Es hat sich jedoch gezeigt, dass die Scheerung im Bild die Erwartungen übertrifft, wenn die reale Aufnahmekonfiguration Kameras dem Modell von Abbildung 24 sehr nahe kommt (vgl. Abbildung 26). Dies ist darauf zurückzuführen, dass schon bei der Berechnung des Trifokaltensors eine projektiv korrekte, geometrisch jedoch verbesserungswürdige Lösung gefunden wurde. Um diese starken Scheerungen zu vermeiden, kann die Qualitätsanalyse der Eckwinkel, die für die Sortierung der Kameras im vorherigen Abschnitt beschrieben wurde, auch für eine Qualitätsanalyse des Trifokaltensors herangezogen werden. Da die Berechnung des Trifokaltensors über robuste Algorithmen wie RANSAC oder die hier verwendete Modifikation GASAC erfolgt, stehen nach der Schätzung in der Regel eine kleinere Menge guter Tensoren zur Verfügung, deren durchschnittlicher Rückprojektionsfehler nur um maximal 0.25 Bildpunkte vom besten Tensor abweicht. Erfüllen mehr als zehn Tensoren dieses Kriterium, werden nur die besten zehn verwendet. Die Tensoren werden anschließend auf ihre Eignung zur Rektifizierung über die Abweichung der rektifizierten Eckwinkel überprüft und der Tensor mit der geringsten Abweichung wird zur Rektifizierung verwendet.

Sind die Rektifizierungstransformationen bestimmt, müssen sie auf die Bilder angewendet werden. Dabei werden in der Regel ganzzahlige Punktrasterkoordinaten auf reelle Zahlen-

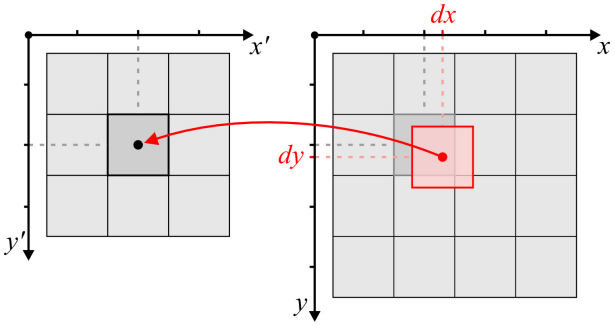


Abbildung 27: Indirekte Transformation

werte transformiert, die wiederum auf ganzzahlige Werte im neuen Punkteraster interpoliert werden müssen. Um zu verhindern, dass dadurch Lücken im Bild entstehen oder Punkte doppelt interpoliert werden, wird wie in der Bildverarbeitung üblich die inverse Transformation angewendet: Dabei wird über die inversen Homographien H^{-1} für jede ganzzahlige Position (i', j') im rektifizierten Bild bestimmt, aus welcher Stelle im Originalbild der Punkt interpoliert werden muss (vgl. Abbildung 27). Die Interpolation erfolgt bikubisch in einer 4x4-Region um die ganzzahligen Punktkoordinaten (i, j) mit dem Subpixelversatz (dx, dy) im Ursprungsbild:

$$[i + dx, j + dy, 1]^T = H^{-1} [i', j', 1]^T$$

$$f(i + dx, j + dy) = \sum_{m=-1}^{2} \sum_{n-1}^{2} f(i + m, j + n) \cdot r(m - dx) \cdot r(dy - n)$$

$$r(k) = \frac{1}{6} \left[p(k + 2)^3 - 4p(k + 1)^3 + 6p(k)^3 - 4p(k - 1)^3 \right] \tag{109}$$

$$p(k) = \begin{cases} k & k > 0 \\ 0 & k \leq 0 \end{cases}$$

Da bei mehreren aufeinander folgenden Interpolationen die Bilder von Mal zu Mal verwaschener wirken, sollte die Anzahl der Interpolationen möglichst gering gehalten werden. Da die Bildpunkte für die Kompensation der radialen Verzeichnung ebenfalls interpoliert werden müssen, bietet es sich an, diese zwei Schritte zu integrieren. Da die radiale Entzerrung jedoch eine nichtlineare Bildtransformation ist, kann man sie in der Regel nicht direkt mit den Homographien verknüpfen und diese Transformationen müssen nacheinander durchgeführt werden. Um trotzdem nur eine Interpolation durchzuführen, wird eine Transformationstabelle generiert, die für jeden Bildpunkt im Zielraster ein Koordinatenpaar gespeichert hat, aus dem hervorgeht von welcher Position im Originalbild interpoliert werden muss. Eine solche Tabelle kann lineare und nichtlineare Transformationsalgorithmen vereinen.

Um diese Tabele zu generieren, berechnet man zunächst aus den radial entzerrten Bildern die Rektifizierungshomographien. Dann wird die Größe der rektifizierten Bilder berechnet, indem die Ecken transformiert werden und für jeden Punkt des rektifizierten Bildes die inverse Homographie angewendet wird. Diese Position bezieht sich auf das gradentreue Bild. Anhand dieser Position und der radialen Verzeichnungsparametern kann nun die endgültige Position im Originalbild berechnet und in der Tabelle eingetragen werden. Jeder Bildpunkt des daraus entstandenen rektifizierten Bildes wurde somit nur einmal interpoliert.

12.2 Ergebnisse

Um Qualität und Eignung der frei gewählten Rektifizierungparameter zu beurteilen, wird der Abstand der Punktkorrespondenzen, die zur Trifokaltensorberechnung verwendet wurden, zu den horizontalen bzw. vertikalen Epipolarlinien untersucht. Das Fehlermaß eines Punkttripels x, x' und x'' ist gegeben durch:

$$
\begin{aligned}
err_{1-2} &= (x^y - x'^y)^2 \\
err_{1-3} &= (x^x - x''^x)^2 \\
err_{2-3} &= ((x^x - x'^x) - (x^y - x''^y))^2/\sqrt{2}
\end{aligned}
\tag{110}
$$

Dieses Fehlermaß umfasst nicht nur den Fehler der Rektifizierung, sondern auch die Fehler bei der Lokalisierung der Korrespondenzen und bei der Trifokaltensorberechnung. Daher soll zusätzlich der Fehler im Sinne der Fehlerfortpflanzung betrachtet werden. So kann angegeben werden, wie stark sich die Fehlerellipsen der Punktmerkmale durch die Rektifizierung verändern. Das Fehlermaß für diese Untersuchung wird über das Verhältnis der Fehlerellipsenachsen vor und nach der Rektifizierung definiert. Bei einem Verhältnis von Eins wird die Fehlerellipse in dieser Achsrichtung nicht vergrößert. Die transformierte Fehlerellipse kann durch die Kovarianzmatrix berechnet werden, die bei der Merkmalsextraktion durch Förstnerpunkte berechnet wurde. Dabei ist zu beachten, dass diese Rektifizierung wieder rückgängig gemacht wird, nachdem die Korrespondenzsuche für alle Bildpositionen (Kapitel 13) durchgeführt wurde. Die Kovarianzen der Homographie interessieren in diesem Fall also nicht, weil sie später wieder invertiert einfließen.

Die Rektifizierungsergebnisse wurden an 272 Bildtripeln untersucht. Zur Beurteilung der Qualität wurden die zur Rektifizierung verwendeten Punkte auf ihre Abweichungen vom Modell untersucht. Jedes Bildpaar hat somit einen mittleren Restfehler nach Gleichung 110 und eine dazugehörige Standardabweichung. In Tabelle 8 sind die minimalen, maximalen und durchschnittlichen mittleren Fehler, daneben die dazugehörigen Standardabweichungen angegeben.

Bildpaar	Restfehler in Punkten			Standardabweichung		
	Min	Max	Ø	Min	Max	Ø
1-2	0.2425	0.3323	0.2816	0.2484	0.3502	0.2929
1-3	0.1989	0.3550	0.2625	0.2032	0.3767	0.2716
2-3	0.2097	0.4177	0.3145	0.2169	0.4852	0.3449

Tabelle 8: Auswertung der Rektifizierung

Der Fehler bleibt deutlich unter einem halben Bildpunkt und die Standardabweichung legt nahe, dass die Werte um den Durchschnittswert normalverteilt sind. Daher wird davon ausgegangen, dass dieser Restfehler zum größten Teil bei der Punktlokalisierung entsteht und nicht durch die Rektifizierung. Weiterhin wird untersucht, inwieweit sich die Fehlerellipsen der Punkte bei der Transformation verändern. Tabelle 9 gibt die Größenveränderung der Fehlerellipsen an, die beim Transfer der Unbestimmtheit der Punkte entsteht. In Abbildung 28 sind die einzelnen Messwerte aus den Tabellen 8 und 9 graphisch aufgetragen. Die x-Achse gibt dabei den Index des Versuchs an. In der ersten und zweiten Reihe gibt die y-Achse den Fehler in Bildpunkten und in der dritten und vierten Reihe das Verhältnis der Fehlerellipsenachsen an.

Bild	Fehlerellipse x-Richtung			Fehlerellipse y-Richtung		
	Min	Max	Ø	Min	Max	Ø
1	0.2530	1.7268	0.9855	0.2683	1.5581	0.9641
2	0.4868	1.6867	0.9953	0.3938	1.5584	0.9724
3	0.3676	1.7394	0.9697	0.3201	1.5275	0.9340

Tabelle 9: Größenveränderung der Fehlerellipsen

Die Minimal- und Maximalwerte in Tabelle 9 zeigen, dass die Fehlerellipsen zwar teilweise verzerrt werden können, dies jedoch sehr selten geschieht (vgl. Abbildung 28 3. und 4. Reihe). Da die durchschnittliche Größenveränderung nahe Eins liegt, ändert die Rektifizierung das Fehlerverhalten nicht stark oder gar nicht.

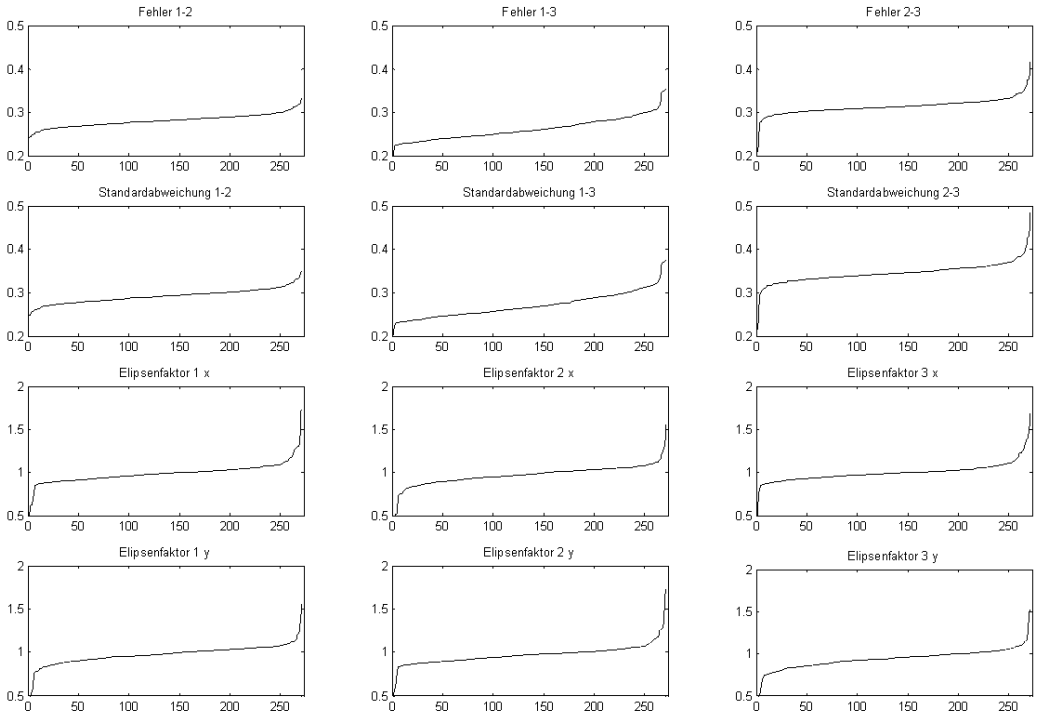


Abbildung 28: Auswertung trinokulare Rektifizierung

13 Räumliche Korrespondenzsuche

Werden zur Korrespondenzsuche nur zwei Bilder verwendet, führt das zu einigen Problemen. Regelmäßige Muster entlang einer Epipolarlinie können nicht eindeutig zugeordnet werden und führen zu fehlerhaften Zuordnungen. Dieser Effekt kommt besonders im Architekturbereich zum Tragen, da hier großflächige, regelmäßige Muster, z.B. Backsteinmauern oder Fensterbänder, sehr häufig vorkommen. Wird die Korrespondenzsuche auf drei Bildern durchgeführt, deren Aufnahmepositionen nicht auf einer Geraden liegen, sind viele Muster nur in einem Bildpaar regelmäßig oder die Frequenz der Muster ist sehr unterschiedlich. Dadurch haben die lokalen Ähnlichkeitsfunktionen stärker ausgeprägte lokale Extrema und werden robuster für Bereiche, die in allen drei Bildern sichtbar sind. Bei Verdeckungen hat die Korrespondenzsuche mit drei Bildern Vor- und Nachteile. Zum einen ist die Zahl der Verdeckungen größer, weil der gemeinsame Sichtbereich kleiner wird, da mehr unterschiedliche Ansichten verwendet werden. Zum anderen können Verdeckungen, die nur in einem Bild auftreten, durch die anderen beiden Bilder rekonstruiert werden, wenn die Punktzuordnungen in diesem Verdeckungsbereich möglich sind. Dadurch wird jedoch die Qualität der Zuordnung in diesem Bereich beeinflusst und es muss zwischen Qualität und Quantität der Punktzuordnungen abgewogen werden.

Ein generelles Problem bei der Korrespondenzsuche liegt in homogenen Bildflächen. Da diese Bildbereiche nicht genügend Informationen aufweisen, um einander eindeutig zugeordnet werden zu können, muss anhand geeigneter Interpolationsverfahren versuchet werden, sie möglichst realitätsnah zu bestimmen. Bei kleinen Bereichen kann hierzu die Fläche durch die umliegenden Kanten interpoliert werden, so dass alle Punkte innerhalb der Fläche einen weichen Übergang zu der begrenzenden Kante aufweisen und keine Doppelzuordnungen auftreten. Bei großen Flächen ist jedoch der Abstand zu den Rändern so groß, dass der Zusammenhang zwischen Kante und Fläche für eine eindeutige Zuordnung nicht mehr ausreicht. Für diese Bereiche bieten sich hierarchische Verfahren an, mit denen die Fläche zuerst in einer geringeren Auflösung grob interpoliert und dann schrittweise verfeinert wird. So sind weiche Übergänge auch in großen homogenen Flächen möglich.

Ein weiteres Problem liegt in der Geschwindigkeit der Korrespondenzsuche. Da die ge-

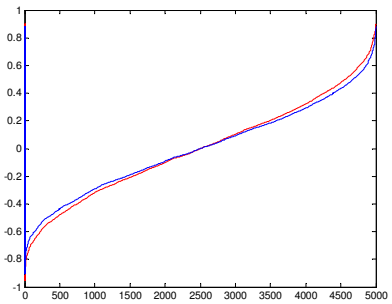


Abbildung 29: Vergleich von MNCC (Blau) zu NCC (Rot)

nerelle Datenmenge kubisch ist (Breite × Höhe × Tiefe), kann kein Algorithmus $O\left(n^3\right)$ unterschreiten, ohne bestimmte Szenenbereiche außer Acht zu lassen. Daher werden vor allem die Algorithmen sehr langsam, die einen noch höheren Aufwand haben. Alleine der Speicherbedarf einer vollständigen Kostenmatrix ist enorm: Ein Bild mit 2048×2048 Punkten und einem Disparitätsbereich von 256 Punkten benötigt bei einer Verwendung von 16 Bit pro Kostenwert bereits 2 GB Speicher. Da 32bit-Applikationen nur maximal 2 GB adressieren können, müssen größere Bilder auf solchen Systemen mit Unterteilungen bearbeitet werden, was im Übergangsbereich dieser Unterteilungen wieder zu neuen Problemen führt. Des weiteren muss schon beim Entwurf des Algorithmus auf Parallelisierbarkeit geachtet werden, da durch die hohe Verfügbarkeit von Mehrkernprozessoren die Skalierbarkeit, also der Geschwindigkeitszuwachs bei mehreren Prozessoren, sehr wichtig ist.

13.1 Lokale Ähnlichkeitsfunktionen

Bei der Wahl der lokalen Ähnlichkeitsfunktionen (vgl. Abschnitt 8.1) müssen mehrere Faktoren berücksichtigt werden. Die verwendete Ähnlichkeitsfunktion sollte robuste und zuverlässige Ähnlichkeitswerte ergeben und invariant sein gegenüber Rauschen, Helligkeits- und Kontraständerung sowie leichten Farbverfälschungen, wie sie durch Fertigungstoleranzen der Kamerasensoren auftreten können. Ferner sollte die Ähnlichkeitsfunktion möglichst lokal sein, weil große Fenster besonders an Objektkanten im Raum zu Ungenauigkeiten führen.

13.1.1 Modifizierte NCC

Die häufig verwendete NCC ist invariant gegenüber Helligkeits- und Kontraständerung, allerdings ist das Rauschverhalten abhängig von der Fenstergröße, was bei verrauschten Bildern zu Problemen an den Kanten führt. Zusätzlich weist dieses Verfahren bei homogenen Flächen eine gravierende Schwäche auf: Der Dividend der Gleichung 18 wird aus der Wurzel des Produktes der Standardabweichung der zu vergleichenden Flächen gebildet. Diese Standardabweichung beträgt bei homogenen Flächen Null, wodurch eine Ähnlickeit mittels NCC nicht berechnet werden kann. Die modifizierte NCC (vgl. Gleichung 19) liefert auch dann noch berechenbare Ergebnisse, wenn nur eine der zu vergleichenden Flächen eine Standardabweichung von Null aufweist. Die Form der Korrelationskurve verändert sich im Vergleich zur NCC geringfügig. Ein nummerischer Vergleich der beiden Funktionen über 5000 zufällige 3×3 -Matrizen ist in Abbildung 29 gegeben. Für beide Verfahren gilt jedoch, dass beispielsweise eine weiße Fläche im Vergleich zu einer schwarzen Fläche eine maximale Ähnlichkeit aufweist, weil durch die Normierung der konstante Farbunterschied wegfällt. Dies gilt für alle homogen gefärbten Flächen.

13.1.2 Punktbasierte Mutual Information

Als Alternative zur Kreuzkorrelation wird in [26] die punktbasierte Mutual Information (MI) vorgeschlagen. Diese Funktion berechnet den Transinformationsgehalt der Farbwerte der Bil-

der, das heißt, wie gut eine Farbe in einem Bilde auf eine Farbe im anderen Bild passt. Diese Farbzuordnung kann beliebig sein. Daher ist die MI invariant gegenüber Intensitäts- und Kontraständerung sowie gegenüber Gammaunterschieden und Farbverfälschungen. Leider birgt sie auch ein prinzipielles Problem: Zu ihrer Berechnung wird eine initiale Korrespondenz der Bildpunkte benötigt, die es zu diesem Zeitpunkt noch gar nicht gibt. Der in [26] vorgeschlagene Ansatz, drei Versuche mit zufälligen Korrespondenzzuordnungen durchzuführen und zu hoffen, dass diese Zufallsfunktionen gegen die wirkliche Funktion konvergieren, birgt den Nachteil, dass die Berechnung nicht deterministisch wird und auch nicht unbedingt zu einem guten Startwert führt. Eine deterministische Initialschätzung ist wegen der Reproduzierbarkeit der Ergebnisse vorzuziehen.

In unserem Fall wird davon ausgegangen, dass die Bilder farblich ungefähr zueinander passen. Ist dies nicht der Fall, kann diese Voraussetzung über eine automatische Histogrammanpassung [69] erzeugt werden. Daher wird für die initiale Korrespondenzsuche in einer stark verkleinerten Auflösung zunächst die Korrespondenzsuche mit der MNCC durchgeführt. Danach wird die MI mit den initialen Korrespondenzen des vorherigen Durchlaufs auf derselben Skala berechnet und die Korrespondenzsuche für diese Skala erneut durchgeführt. Der zusätzliche Rechenaufwand ist in diesem Fall sehr gering, weil die Auflösung in der kleinsten Skala 128×128 Punkte nicht stark überschreitet. Passen die Bilder farblich exakt zueinander, beispielsweise bedingt durch die Verwendung derselben Kamera mit denselben Einstellungen für Belichtung und Weißabgleich, kann auch die identische Funktion als initiale Schätzung verwendet werden. Allerdings sollte diese Funktion aus statistischen Gründen durch eine Gaußfaltung etwas gespreizt werden. Bei der Verarbeitung von Videosequenzen ändert sich der Transinformationsgehalt von einem Bild zum nächsten nur sehr geringfügig. Daher ist es in einem solchen Fall auch möglich, die MI des letzten Bildes für das nächste wiederzuverwenden. Die geringfügigen Veränderungen durch die unterschiedliche Beleuchtung und unterschiedlichen Objekte im Bild können berücksichtigt werden, indem man nach erfolgreicher Korrespondenzsuche die MI für das aktuelle Bild neu berechnet und sie dann wiederum an das nächste weitergibt.

Eine Schwäche der MI liegt in der Beschränkung auf Grauwerte, die dadurch zu begründen ist, dass bei 24bit-Farbbildern eine sehr große Wertemenge statistisch auszuwerten ist $\left(2^{24} = 16777216\right)$. Dadurch sind die gemischten Wahrscheinlichkeiten (Gleichung 23) selbst bei Bildern mit 10 Millionen Bildpunkten nicht repräsentativ und die Korrelationstabelle für alle Farben würde ebenfalls zu groß $(2^{24} \times 2^{24})$. Ferner sind die einzelnen Farbkanäle bei RGB nicht voneinander unabhängig. Dies könnte zwar durch eine andere Farbrepräsentation (Lab, HSV oder LUV) geändert werden, wodurch allerdings das prinzipielle Problem des großen Wertebereich nicht umgangen werden kann.

Daher wird ein pragmatischer Ansatz gewählt und die RGB-Bilddaten werden unabhängig voneinander betrachtet. Dazu wird mit Gleichung 23 für jede Punktkorrespondenz eine separate Wahrscheinlichkeit für Rot, Grün und Blau generiert und damit drei MI-Kostentabellen nach Gleichung 25 mit separater Normierung nach Gleichung 26 erstellt. Jede Bildkorrespondenz hat somit drei MI-Korrelationswerte. Die lokale Ähnlichkeitsfunktion für Farbbilder ergibt sich aus dem arithmetischen Mittel dieser drei Werte.

13.2 Semi-global Matching und Erweiterungen

Die Korrespondenzsuche wird prinzipiell mittels SGM durchgeführt. Daher werden in Abschnitt 13.2.1 die wesentlichen Aspekte dieses Verfahrens ausfürlich beschrieben. Die darauffolgenden Abschnitte behandeln die vorgeschlagenen Erweiterungen des SGM, die im Rahmen dieses Systems entwickelt wurden. Diese bestehen hauptsächlich aus der Anpassung des Algorithmus an den Dreibildfall, der optionalen Minimierung des Speicherplatzbedarfs von $O(n^3)$ auf $O(n^2)$ und der Verwendung hierarchischer Tiefeninformationen. Detailverbesserungen umfassen die Verwendung von Gradientenkarten und ein verbessertes Verfahren zur Medianfilterung der Ergebnisse. Des weiteren wird die Parallelisierbarkeit des Algorithmus untersucht und durchgeführt. Eine ausführliche Analyse eines Großteils dieser Techniken wurde in [23] veröffentlicht.

13.2.1 Semi-global Matching

Beim SGM wird versucht, die Güte der Korrespondenzsuche über eine Energiefunktion der Disparitätenkarte zu modellieren:

$$E(D) = \begin{array}{l} \sum\limits_{x,y\epsilon I} \left(\left(1-\rho\left(a\left(x,y\right),b\left(x+D\left(x,y\right),y\right)\right)\right)+\right. \\ Q_1 \sum_{i,j=-1}^{1} T\left[\left|D\left(x,y\right)-D\left(x+i,y+j\right)\right|=1\right]+ \\ \left. Q_2 \sum_{i,j=-1}^{1} T\left[\left|D\left(x,y\right)-D\left(x+i,y+j\right)\right|>1\right]\right) for\ i \neq j \end{array} \tag{111}$$

Die Energiefunktion besteht aus drei Teilen: Der erste Teil bestimmt die lokalen Kosten aus dem lokalen Ähnlichkeitsmaß ρ des zu untersuchenden Punktepaares oder -tripels. Der zweite Teil versucht, die Disparitäten auf derselben Ebene orthogonal zur Bildebene zu halten und bestraft Disparitätssprünge um Eins mit einem niedrigen Kostenfaktor Q_1. Dieses Verhalten ermöglicht es, auch leicht schräg zur Bildebene liegende Ebenen im Raum zu rekonstruieren. Der letzte Teil mit Kostenfaktor Q_2 ist für größere Disparitätssprünge vorgesehen. Sind die lokalen Kosten so hoch, dass ein Disparitätssprung energetisch sinnvoller ist als ein Verbleiben auf der Ebene, werden die Kosten an dieser Stelle mit einem konstanten Kostenfaktor belegt. Somit wird an einer solchen Stelle der Sprung bevorzugt, der die niedrigsten lokalen Kosten verursacht. Da eine vollständige Analyse sämtlicher Disparitätskarten und Sprungmöglichkeiten exponentiellen Aufwand bedeuten würde und eine Baumsuche nach den geringsten Kosten ein NP-hartes Problem darstellt, kann diese Energiefunktion für größere Bilder mit vertretbarem Aufwand nur näherungsweise bestimmt werden. Dazu wird die lokale Kostenmatrix für alle Korrespondenzkandidaten über alle Disparitäten $d\epsilon\left[d_{min}, d_{max}\right]$ erzeugt und aus verschiedenen Richtungen linear approximiert, was einen algorithmisch günstigen Effekt mit sich bringt: Jeder Punkt hängt pro Richtung nur von den Disparitätswerten seines direkten Vorgängers und dem geringsten Kostenwert L_{min}^{i+1} aus dem vorherigen Iterationsschritt ab. Daher kann diese Analyse rekursiv über $i\epsilon\left\{n\cdots 1\right\}$ und sehr speichersparend implementiert werden:

$$\begin{array}{rl} L^n\left(x_n,y_n,d\right) = & 1-\rho\left(a\left(x_n,y_n\right),b\left(x_n+d,y_n\right)\right) \\ L^i\left(x_i,y_i,d\right) = & 1-\rho\left(a\left(x_i,y_i\right),b\left(x_i+d,y_i\right)\right)+ \\ & \min(L^{i+1}\left(x_{i+1},y_{i+1},d-1\right)+Q_1, L^{i+1}\left(x_{i+1},y_{i+1},d\right), \\ & L^{i+1}\left(x_{i+1},y_{i+1},d+1\right)+Q_1, L_{min}^{i+1}+Q_2) \\ L_{min}^{i+1} = & \min\limits_{d\epsilon[d_{min},d_{max}]} L^{i+1}\left(x_{i+1},y_{i+1},d\right) \end{array} \tag{112}$$

Die Richtungen können über Richtungsvektoren r_j definiert werden. Die Punktpositionen ergeben sich dann aus:

$$\begin{array}{c} r_j = \begin{bmatrix} r_j^x \\ r_j^y \end{bmatrix} \\ x_i = x_1 + (i-1)\cdot r_j^x \\ y_i = y_1 + (i-1)\cdot r_j^y \end{array} \tag{113}$$

Die Anzahl der Richtungen sollte mindestens Acht, entsprechend den vier Hauptrichtungen und den vier Diagonalen (vgl. Abbildung 30) betragen.

Um Überläufe des Wertes im Speicherregister durch das rekursive Aufsummieren zu verhindern, wird in jeder Iteration der Minimalwert L_{min}^{i+1} der letzten Iteration abgezogen, wodurch sich die Charakteristik der Kostenfunktion jedoch nicht ändert. Die Disparitätenkarte lässt dich nun aus dem Minimum der Summe aller Richtungen r_j bestimmen:

$$D\left(x,y\right) = \min_{d}\left(\sum_{j=1}^{8} L_{r_j}\left(x,y,d\right)\right) \tag{114}$$

$$r_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} : \longrightarrow \qquad r_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} : \longleftarrow \qquad r_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} : \downarrow$$

$$r_4 = \begin{bmatrix} 0 \\ -1 \end{bmatrix} : \uparrow \qquad r_5 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} : \searrow \qquad r_6 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} : \swarrow$$

$$r_7 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} : \nearrow \qquad r_8 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} : \nwarrow$$

Abbildung 30: Richtungen für die SGM-Approximation

Um stabilere Korrespondenzen zu erhalten, wird eine Symmetrieprüfung durchgeführt: Die zu registrierenden Bilder werden miteinander vertauscht und nur die Korrespondenzen als gültig erachtet, mit denen in beiden Fällen der selbe Korrespondenzpartner ermittelt wurde.

13.2.2 Dreibildfall

Die Korrespondenzsuche auf drei Bildern, die mit dem in Kapitel 12 beschriebenen Verfahren rektifiziert wurden, entspricht im Grundsatz der Suche auf zwei Bildern. Da sich die Disparitäten des sekundären und des tertiären Bildes entsprechen, können die zwei Ergebnisse der lokalen Ähnlichkeitsfunktion eindeutig zugeordnet und gemittelt werden. Das Ergebnis wird daraufhin in eine gemeinsame Kostenmatrix eingetragen. Nachdem diese Matrix für alle in Frage kommenden Disparitäten an allen Bildpositionen bestimmt wurde, wird das SGM für den Zwei- und Dreibildfall wieder auf dieselbe Art und Weise berechnet, abgesehen von der nahe liegenden Erweiterung, dass für die symmetrische Korrespondenzsuche im Dreibildfall natürlich die Disparitätenkarten aller drei Bilder berechnet werden. Durch die Mittelung der Ähnlichkeitsfunktion wird der lokale Einfluss besser bestimmt und die Robustheit erhöht, weil Symmetrieeffekte entlang einer Epipolarlinie reduziert werden, was insgesamt zu besseren Ergebnissen führt.

In Bereichen, in denen die aktuelle Disparität nur in einem Bild sichtbar ist, muss entschieden werden, ob auf Kosten der Dichte ganz auf ein Ergebnis verzichtet oder ob in diesem Bereich auf Kosten der Robustheit auf das klassische Zweibildverfahren zurückgegriffen werden soll. Eine ähnliche Entscheidung muss bei der Bestimmung der symmetrischen Disparitätenkarten gefällt werden. Ist die Disparität in zwei Bildern symmetrisch, im dritten aber nicht, muss entschieden werden, ob dieses teilweise inkonsistente Ergebnis, das in der Regel durch Verdeckungen hervorgerufen wird, verworfen werden soll. Die andere Möglichkeit besteht darin, die inkonsistenten Daten entsprechend der anderen Bilder zu korrigieren. So können sogar Bereiche, die nur in einem Bild verdeckt sind, erfolgreich zugeordnet werden.

13.2.3 Hierarchisches Modell

Bei der Verwendung eines hierarchischen Modells wird das Bild zunächst verkleinert und die Korrespondenzsuche nacheinander auf immer größeren Skalierungen des Bildes durchgeführt, bis die ursprüngliche Bildgröße wieder erreicht ist. In [26] wird zwar empfohlen, die Mutual Information hierarchisch zu berechnen, die Zwischenergebnisse aufgrund der potenziellen Fehlerfortpflanzung jedoch nicht zu verwenden. Diese Argumentation berücksichtigt allerdings nicht den Aspekt, dass ein hierarchischer Ansatz wertvolle Zusatzinformationen liefert und damit den Informationsmangel in homogenen Gebieten teilweise behebt. Zusätzlich kommt es vor, dass in Gebieten mit guter, aber regelmäßiger Textur viele Fehlzuordnungen ermittelt werden, wenn die Disparität nicht durch einen hierarchischen Ansatz eingeschränkt wird.

Ein weiterer Aspekt betrifft die Reduzierung des Suchraumes. Da bei einer vorhandenen Tiefenkarte nur noch ein reduzierter Bereich abgesucht werden muss, verringert sich der Rechenaufwand enorm. Um den Suchbereich korrekt einzuschränken, müssen die Kandidaten mehrere Kriterien erfüllen:

1. Kandidaten können nur dann ausgeschlossen werden, wenn sie von anderen Korrespondenzen belegt sind. Dazu wird geprüft, ob benachbarte Bildpunkte bereits Korrespondenzen aus dem aktuellen Suchbereich des aktuellen Punktes belegt haben.

2. Befindet sich der aktuelle Punkt an einem Disparitätssprung, weisen die beiden direkten Nachbarpositionen entlang der Epipolargeraden also eine Disparitätsdifferenz größer Eins auf, sollte der gesamte Bereich zwischen diesen Nachbarn untersucht werden, da sich die Korrespondenz sowohl an der unteren Kante als auch im Zwischenraum oder auf der oberen Kante der Disparitätsstufen befinden kann. Dieser Bereich wird direkter Suchbereich genannt.

3. Ist nur einem der direkten Nachbarn eine korrekt bestimmte Korrespondenz zugeordnet, ist der Suchraum nur in dieser Richtung begrenzt und muss in die andere Richtung vollständig geöffnet werden.

4. Um die Vergrößerung der Auflösung zu berücksichtigen, fügt man dem ermittelten direkten Suchbereich auf beiden Seiten einen konstanten Bereich hinzu.

Die Reduzierung des Suchraumes wird in Abbildung 31 verdeutlicht. Auf der unteren Seite ist ein Streifen des Referenzbildes gezeigt, dessen Bildpunkte mit den vertikal aufgetragenen Streifen des Sekundärbildes auf der linken Seite verknüpft werden müssen. An jeder Position ist der Suchraum aufgetragen, woraus sich der diagonale Bildbereich ergibt. Der schwarze Bereich kennzeichnet den Suchbereich, der nicht untersucht werden muss, und die grauen Bereiche die Werte, die als mögliche Kandidaten untersucht wurden. Die rote Linie in der diagonalen Fläche zeigt die gefundene Korrespondenz für die Punkte auf der roten Linie in den Bildausschnitten nach der SGM-Berechnung. Ein Beispiel für den weiß markieten Punkt im Referenzbild ist über den grauen Linien und dem weiß umrandeten Kasten im Sekundärbid verdeutlicht. Durch den Einsatz der Hierarchie konnten über 60% der möglichen Korrespondenzen bereits vor der Berechnung ausgeschlossen werden.

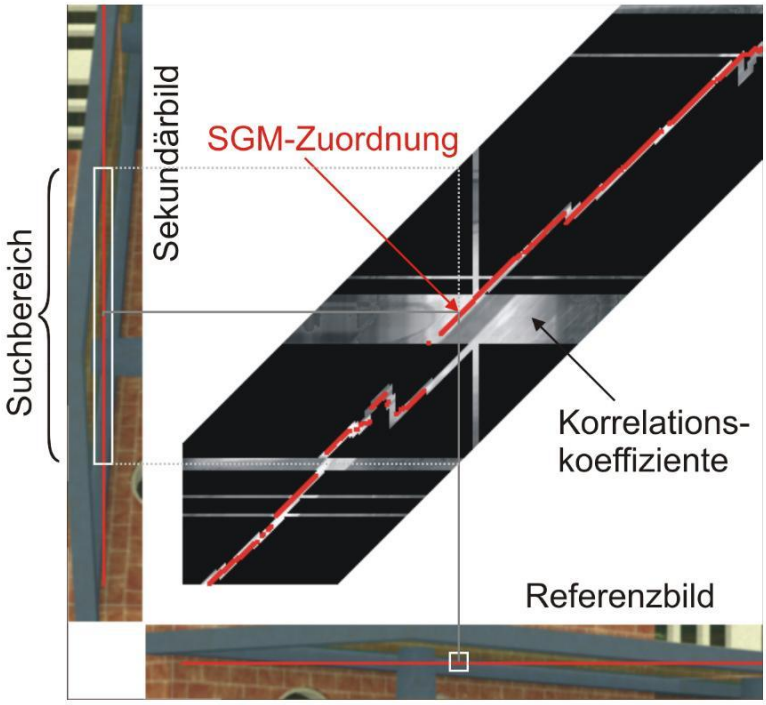


Abbildung 31: Suchraum mit hierarchischem Modell

Die Zahl der Hierarchiestufen ist abhängig von der Bildauflösung und dem Skalierungsschritt. Der Skalierungsschritt wird üblicherweise auf Zwei gesetzt, die Bildgröße der kleineren Seite sollte in der kleinsten Auflösung eine gewisse Mindestgröße nicht unterschreiten.

Als Minimalgröße wird 128 vorgeschlagen, woraus sich für die Anzahl der Stufen l folgende Berechnung ergibt:

$$l = 1 + Rundung\left(log_2\left(\frac{min\,(w,h)}{128}\right)\right) \tag{115}$$

13.2.4 Speicheroptimierung durch Pfadlängenbegrenzung

Die Energiefunktion 111 hat die besondere Eigenschaft, dass nach einem Tiefensprung der Kontext der vorangegangenen Punkte aus dieser Richtung keinen Einfluss mehr auf die Ergebnisse der folgenden Punkte hat. Dadurch ist der Einfluss entfernter Punkte in der Regel begrenzt und die in Gleichung 112 durch i definierte Pfadlänge kann mit geringem Genauigkeitsverlust auf einen festen Wert $e\epsilon\,[3,10]$ begrenzt werden [23]. Hierbei ist zu beachten, dass die Pfadlänge für die beiden horizontalen Richtungen nicht begrenzt werden muss, da die Ergebnisse direkt in der Zeile akkumuliert werden können. Die Größe der Kostenmatrix wird dadurch von $O\,(n^3) = (w \cdot h \cdot d)$ auf $O\,(n^2) = (w \cdot 2e \cdot d)$ verringert.

Da sich der Einflussbereich allerdings für jede neue Zeile um eine Zeile verschiebt, muss die Kostenmatrix für jede Zeile neu aufgestellt werden. Dadurch vergrößert sich der Rechenaufwand genau um den konstanten Faktor $2e$. Aus diesem Grund wurde die Entwicklung der Algorithmen auf 64bit-Betriebssysteme umgestellt, wodurch die 4GB-Grenze für Applikationen wegfällt und die speicherintensivere, aber schnellere Variante des SGM ohne Pfadlängenbegrenzung verwendet werden kann.

13.2.5 Gradienten: Non-Maxima-Unterdrückung

Der Kostenfaktor Q_2 für große Tiefensprünge in Gleichung 112 behandelt homogene Flächen und Bildkanten gleich. Besonders an den Bildkanten ist jedoch ein Tiefensprung wahrscheinlicher als in homogenen Flächen. Daher wird vorgeschlagen, diesen Kostenfaktor abhängig von der Gradientenstärke zu machen und den Wert Q_2 an Positionen mit starken Gradienten an den Wert Q_1 anzunähern. Um zu verhindern, dass Kanten, die über mehrere Punkte laufen, zu unscharfen Kostengrenzen führen, wird das Kantenbild zunächst mit einer Non-Maxima-Unterdrückung in Kantenrichtung gefiltert. Dieses Verfahren der Non-Maxima-Unterdrückung entspricht der Canny-Edge-Filterung ohne den abschließenden Hystereseteil [6]. Durch den Verzicht auf den Hystereseteil wird vermieden, dass zwei weitere Parameter für die Kantenfilterung benötigt werden. Außerdem führen schwache Kanten so zu einer geringeren Annäherung von Q_2 an Q_1. Das Kantenbild wird aus der Gradientenstärke der Gaußrichtungsableitungen g_x und g_y berechnet. Das Sigma der Gaußglättung für die Richtungsableitungen wurde auf $\sigma = 0.85$ gesetzt, da dieser Wert dem Sobeloperator recht nahe kommt. Um das Bildrauschen zu unterdrücken, wurden alle so berechneten Kanten, die unterhalb von 85% des maximalen Gradienten liegen auf Null gesetzt und der Gradient auf das Intervall $[0,1]$ normiert. Der lokale, gewichtete Kostenfaktor ergibt sich aus:

$$\begin{aligned} grad\,(x,y) &= \sqrt{\left(g_x\,(x,y)\right)^2 + \left(g_y\,(x,y)\right)^2} \\ max_{grad} &= \overset{\max}{x,y}\,(grad\,(x,y)) \\ gradQ_2\,(x,y) &= T\left[grad\,(x,y) \geq 0.85 \cdot max_{grad}\right] \cdot nonMax\,(grad\,(x,y)) \\ Q_2\,(x,y) &= Q_2\left(1 - \tfrac{gradQ_2(x,y)}{max_{grad}}\right) \end{aligned} \tag{116}$$

13.2.6 Medianfilterung

Nach der Symmetrieprüfung weisen die Bilder häufig 1-4 Bildpunkte große Löcher in der Disparitätenkarte auf, die durch eine Medianfilterung mit einer Maskengröße von 5×5 wieder geschlossen werden können. Hierbei dürfen nur Stellen aufgefüllt werden, die nach

der Symmetrieprüfung selbst keine Korrespondenz mehr haben, jedoch von mindestens 13 Nachbarn mit gültigen Korrespondenzen umgeben sind. Die Tiefenkarte kann zwar danach mit einem weiteren Medianfilter geglättet werden, allerdings sollte die Maskengröße dieses Filters immer kleiner sein als die Maskengröße für das Füllen von Fehlstellen.

13.2.7 Parallelisierbarkeit

Bei modernen Multiprozessorrechnern stellt sich in immer höherem Maße die Frage nach der Parallelisierbarkeit der Algorithmen. Besonders die Korrespondenzsuche per SGM ist für die Parallelisierbarkeit ein guter Kandidat. Der Algorithmus lässt sich abgesehen von Ein- und Ausgabe in fünf wesentliche Schritte unterteilen:

1. Berechnung der Mutual Information
2. Aufstellen der lokalen Kostenmatrix
3. Akkumulieren der SGM-Pfade
4. Bestimmung der minimalen Kosten
5. Filtern der Disparitätenkarten mittels
   (a) symmetrischer Korrespondenzsuche
   (b) Medianfüllung
   (c) Medianfilterung

Die Mutual Information braucht zur Berechnung die gemischten Korrespondenzen aus der Disparitätenkarte des vorherigen Durchlaufes. Jedem verfügbaren CPU kann ein kleiner Ausschnitt der Disparitätskarte zugeordet werden, die er untersucht. Die Berechnung der 256×256-Lookuptables für die Farbkanäle kann in konstanter Zeit erfolgen.

Für die berechnung der lokalen Kosten werden die Bilder in gleich große horizontale Segmente untergliedert und auf die verfügbaren CPUs verteilt. Dieser Teil des Algorithmus skaliert linear mit der Anzahl der CPUs, bis die Bildzeilenzahl erreicht ist.

Das Akkumulieren der SGM-Pfade wird parallelisiert, indem zuerst die einzelnen Richtungen aus Gleichung 114 auf jeweils eine separate CPU verteilt werden. Stehen mehr CPUs ans Richtungen zur Verfügung, können die Pfade entlang ihrer Richtung analog zur lokalen Kostenmatrix in separate Segmente unterteilt werden. Da die Pfadkosten einer Richtung auf eine gemeinsame, globale Kostenmatrix addiert werden, kann es zu sogenannten *race conditions* kommen, wenn zwei CPUs auf exakt demselben Wert der globalen Kostenmatrix arbeiten. An dieser Stelle sind die CPUs von einander abhänig und müssen warten, bis eine CPU mit der Berechnung fertig ist, um das Ergebnis der anderen zur Verfügung zu stellen. Das Überprüfen dieser *race conditions* limitiert aber nicht die Parallelisierbarkeit und tritt ohnehin nur an maximal an n^2 Stellen der globalen Kostenmatrix auf, an denen sich die Pfade kreuzen, wobei n die Anzahl der CPUs ist.

Die Bestimmung der minimalen Kosten sowie die Symmetrieprüfung der Korrespondenzen kann für jede Bildposition unabhängig von den anderen erfolgen. Auch die Medianfilterung kann auf verschiedene Bildsegmente verteilt werden, wenn die Eingangsdaten zunächst nicht verändert werden und das Endergebnis erst zum Schluss mit den Eingangsdaten vertauscht wird.

Daher skaliert die Korrespondenzsuche mittels SGM mit der CPU-Anzahl bis zur Anzahl der Bildzeilen bzw. -spalten linear. Diese Begrenzung auf die Anzahl der Bildzeilen ergibt sich aus Schritt 3. In der Regel stehen jedoch weit weniger CPUs als Bildzeilen zur Verfügung, weshalb diese Limitierung in der Praxis noch nicht relevant ist.

13.3 Subpixelberechnung

Die Subpixelberechnung der Korrespondenzposition erfolgt prinzipiell nach dem in [15] vorgestellten iterativen Verfahren. Dazu wird für jede Korrespondenz des Bildes die Subpixelposition entlang der Epipolarlinie berechnet, indem zunächst der SSD-Wert des Punktpaares $(x, y(d))$ mit der ganzzahligen Disparität d sowie die SSD-Werte der beiden Nachbarn $(x, y(d-1))$ und $(x, y(d+1))$ berechnet werden:

$$\begin{aligned} s_1 &= \varrho_{SSD}(x, y(d-1)) \\ s_2 &= \varrho_{SSD}(x, y(d)) \\ s_3 &= \varrho_{SSD}(x, y(d+1)) \end{aligned} \tag{117}$$

Daraufhin wird eine Parabel f bestimmt, die durch diese drei SSD-Werte definiert ist. Die Subpixelposition d_{sub} ergibt sich nun aus der Extremstelle ext von f. Da die Abstände der SSD-Werte auf der x-Achse gleich Eins sind und um Null liegen, kann die Berechnung der Extremstelle zu folgender Gleichung umgeformt werden:

$$\begin{aligned} f(x) &= a^2 x + bx + c \\ a &= \tfrac{s_3 - s_1}{2} - (s_2 - s_1) \\ b &= (s_2 - s_1) - a \\ c &= s_1 + b - a \\ ext(f) &= \tfrac{-b}{2a} \\ d_{sub} &= d + ext(f) \end{aligned} \tag{118}$$

Aus dem Absolutwert der Krümmung a der Parabel ist ersichtlich, wie gut diese Position approximiert wurde: Große Krümmungswerte deuten auf ein scharfes lokales Maximum hin, während leichte Krümmungswerte auf eine ungenaue Subpixelapproximation schließen lassen. Im vorliegenden Dreibildfall muss zusätzlich entschieden werden, welches Bildpaar für die Subpixelbestimmung verwendet werden soll. Eine Akkumulation der Subpixelwerte ist hier nicht sinnvoll, da die Alias-Effekte, die bei der Subpixelbestimmung ausgenutzt werden, im horizontalen und vertikalen Bildpaar nicht gleich ausgeprägt sind. Daher wird die Subpixelbestimmung für beide Paare unabhängig voneinander durchgeführt und danach der Wert übernommen, der eine stärker gekrümmte Parabel hat und damit die größere Struktur im Bild aufweist.

Da diese Subpixelposition stark verrauscht ist, wird versucht, eine Energiefunktion über alle Disparitäten d zu approximieren, die dieses Rauschen abhänig von der lokalen Varianz und der Parabelstärke glättet. Dazu werden die Werte der Parabelkrümmung auf das Intervall $[0;1]$ normiert. Die Energiefunktion besteht aus zwei Teilen, der lokalen Subpixelposition d_{sub} gewichtet mit der lokalen Parabelstärke a und einem Glättungsteil. Dieser Glättungsteil ist so modelliert, dass er den Durchschnittswert der Disparitäten $\overline{d}$ und die Varianz σ in einer kleinen Umgebung N um die Disparität d_0 berechnet und mit einem angenommenen Sollwert σ_{norm} vergleicht. Aus diesem Glättungswert und der lokalen gewichteten Subpixelposition ergibt sich die neue Subpixelposition d_0:

$$\begin{aligned} \overline{d} &= \tfrac{1}{N} \textstyle\sum_{i=0}^{N-1} d_i \\ \sigma^2(d_0) &= \tfrac{1}{N-1} \textstyle\sum_{i=0}^{N-1} \left(d_i - \overline{d}\right)^2 \\ H &= \tfrac{\lambda_{sub}}{N} \tfrac{1}{1+(\sigma/\sigma_{norm})^4} \\ d_0 &= \tfrac{a \cdot d_{sub} + H\overline{d}}{a+H} \end{aligned} \tag{119}$$

Wobei λ_{sub} ein konstanter globaler Gewichtungsfaktor zwischen Glättung und lokaler Stärke ist. Die Herleitung dieser Berechnung ist in [15] gezeigt.

Um den Glättungseffekt der Disparitäten auf größere Bereiche auszudehnen, wird diese Adaption der Subpixelverschiebung iterativ durchgeführt, bis sich die Disparitätswerte nicht mehr signifikant vom letzten Iterationsschritt unterscheiden.

Das originale Verfahren glättet jedoch über Kanten hinweg, weil die ganzzahligen Disparitätsunterschiede bei der Varianzberechnung nicht berücksichtigt werden. Daher wird diese Technik um eine Disparitätsprüfung erweitert. Für die Mittelwert- und Varianzberechnung werden nur Disparitäten verwendet, die unterhalb des gegebenen Schwellwerts von Eins liegen. Die Berechnung der Varianz σ und die Anzahl der Punkte in der Umgebung N werden daher durch folgende Gleichung ersetzt:

$$\begin{aligned} N_{adapt} &= \sum_{i=0}^{N-1} T\left[|d_i - d_0| \leq 1\right] \\ \sigma^2\left(d_0\right) &= \frac{1}{N_{adapt}-1} \sum_{i=0}^{N-1} T\left[|d_i - d_0| \leq 1\right] \cdot \left(d_i - \overline{d}\right)^2 \end{aligned} \tag{120}$$

Die Subpixelapproximation glättet somit nicht mehr über starke Kanten der Disparitätenkarte hinweg, wird aber ausgesetzt, wenn die Anzahl der unterstützenden Punkte N_{adapt} unter $N/2$ fällt.

13.4 Zusammenfassung des Algorithmus

Die erklärten Techniken können nun zu Algorithmus 5 zusammengefasst werden:

Algorithmus 5 Hierarchisches Semi-global Matching

1. Verkleinere die Bilder in 2-Potenzschritten, bis die kleinere Seite noch größer als 128 ist.
2. Für alle Bilder:
   (a) Berechne eine lokale Kostenmatrix für alle Disparitäten entlang der Epipolarlinien mittels MNCC.
   (b) Bei drei Bildern: Bilde das arithmetische Mittel der zwei Kostenmatrizen.
   (c) Berechne die Kostenmatrix für Q_2 anhand der Gradienten.
   (d) Berechne die SGM aus acht Richtungen.
   (e) Bestimme die Disparität anhand der minimalen Kosten.
3. Prüfe, ob die Punkte symmetrische Zuordnungen in den Bildern haben.
4. Führe eine Medianreparatur mit 5×5-Maske durch.
5. Führe eine Medianglättung mit 3×3-Maske durch.
6. Für alle Skalen:
   (a) Skaliere die letzten Disparitäten (außer für die erste Skala).
   (b) Berechne anhand dieser Disparitäten die MI.
   (c) Für alle Bilder:
      i. Berechne eine lokale Kostenmatrix für alle Disparitäten entlang der Epipolarlinien mittels MI.
      ii. Bei drei Bildern: Bilde das arithmetische Mittel der zwei Kostenmatrizen.
      iii. Berechne die Kostenmatrix für Q_2 anhand der Gradienten.
      iv. Berechne die SGM aus acht Richtungen.
      v. Bestimme die Disparität anhand der minimalen Kosten.
   (d) Prüfe, ob die Punkte symmetrische Zuordnungen in den Bildern haben.
   (e) Führe eine Medianreparatur mit 5×5-Maske durch.
   (f) Führe eine Medianglättung mit 3×3-Maske durch.
7. Berechne die Subpixelverschiebung.

13.5 Ergebnisse

Die vorgestellten Techniken werden anhand Experimenten an eigenen, synthetischen Daten und an den Bildern der Middlebury Stereo Vision Hompage analysiert. In diesem Abschnitt wird zunächst die Kostenberechnung der Mutual Information analysiert. Im Anschluss werden anhand eines eigenen synthetischen Datensatzes mit bekannten Verschiebungen Vergleiche der einzelnen Kostenfunktionen, der Einfluss von hierarchischer Korrespondenzsuche, Dreibildfall, Medianfilterung und Subpixelberechnung sowie der Glättungsparameter untersucht. Zuletzt werden die Ergebnise mit Hilfe der Middlebury-Daten mit anderen bekannten Stereoverfahren verglichen.

13.5.1 Analyse der Mutual Information

Für die MI werden synthetische Bilder mit synthetischer Verschiebung analysiert. Die Farben im Referenzbild sind gleichverteilt und das Sekundärbild wurde anhand einer gegebenen Verschiebungsmatrix generiert. Danach wurden Helligkeit, Kontrast, Gamma, Rauschverhalten

und Farbton im Sekundärbild verändert und die MI über die bekannte Verschiebungsmatrix bestimmt. Auf diese Weise ist die echte MI des Bildes ermittelt und kann mit der über SGM bestimmten MI verglichen werden. Um die Qualität der MI zu beurteilen, wird die Kostentabelle der MI als Grauwertbild dargestellt und ist in Abbildung 32 zu sehen. Bei der Farbverfälschung wurde der rote Kanal anders verfälscht als der grüne und der blaue. Bei den restlichen Versuchen war die MI für alle Kanäle gleich.

Für eine quantitative Auswertung werden die Diagramme in Abbildung 32 auf ihre durchschnittliche prozentuale Abweichung analysiert und die Standardabweichung wird angegeben (Tabelle 10). Die sehr konstante Standardabweichung im Bereich von 2.07 - 2.82% zeigt, dass die vorgeschlagene Kostenfunktion gegenüber komplexen und nicht linearen Farbänderungen sehr stabil ist. Die Zunahme der durchschnittlichen Abweichung liegt selbst bei Farbbereichsänderungen und Rauschen unter 2.33%.

	Orig.	Hell.	Kontr.	Gamma	R. 10	R. 25	Farbe
Ø	2.45%	4.78%	3.84%	3.64%	2.98%	3.89%	2.72%
σ	2.58%	2.45%	2.07%	2.01%	2.82%	2.75%	2.23%

Tabelle 10: Nummerische Analyse der Mutual Information

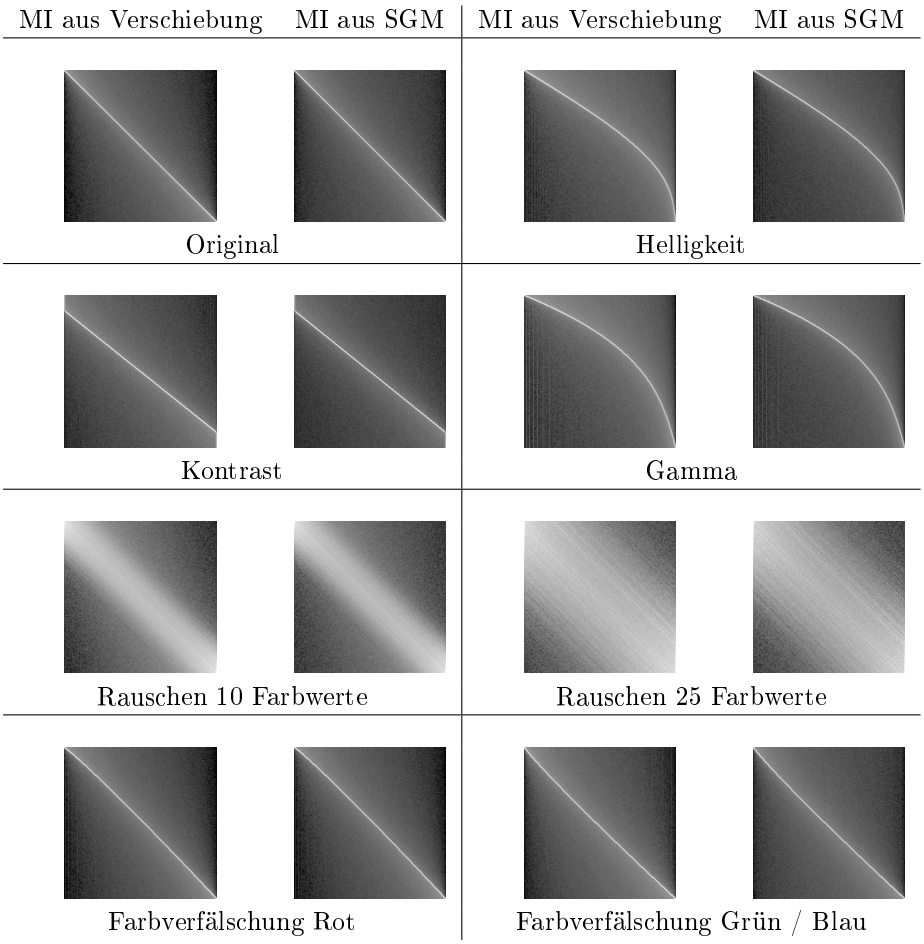
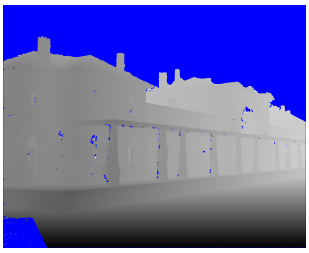


Abbildung 32: Graphische Analyse der Mutual Information

13.5.2 Quantitative Analyse des SGM mit synthetischen Daten

Zur Analyse der unterschiedlichen Parameter und Techniken des SGM wird ein synthetisches Bildtripel der Größe 1280×1024 aus einem 3D-Modell generiert. Von diesem Tripel sind

Referenzbild



Disparitätenkarte

Abbildung 33: Synthetische Bilder für die SGM-Analyse

die Disparitäten bekannt und werden mit dem Ergebnis des SGM verglichen. Dabei wird untersucht, wieviele korrespondierende Punkte gefunden wurden und wieviele davon korrekt sind, also einen Fehler geringer ist als Eins aufweisen. Ein Problem bei dieser Analyse ist die Klassifizierung von Ausreißern. Diese sind nicht gaußverteilt, sondern steigen in ihren Werten sprunghaft an. Die Angabe von durchschnittlichem Fehler und Standardabweichung ist daher für den ganzen Datensatz nicht sinnvoll. Sie werden daher nur über die Punkte ermittelt, deren absoluter Fehler kleiner ist als Zwei. Die synthetischen Bilder für die Analyse sind in Abbildung 33 gezeigt. Links befindet sich eines der drei Referenzbilder. Daneben ist beispielhaft eine berechnete Disparitätenkarte gezeigt. Der Himmel wurde ausmaskiert, da er als Textur ohne korrekte 3D-Information den oberen Bildbereich ausfüllt.

Die Ergebnisse der durchgeführten Experimente sind in Tabellen 12 und 14 zu sehen. In der ersten Spalte ist eine Beschreibung der modifizierten Parameter beziehungsweise der Wert des Parameters angegeben, in der zweiten Spalte die prozentuale Erkennungsrate der korrekt zugeordneten Punkte. Die folgenden beiden Spalten führen die durchschnittlichen Fehler der korrekt zugeordneten Punkte sowie die Standardabweichungen in Bildpunkten auf. In der letzten Spalte stehen die Laufzeiten der Experimente in Sekunden. Der beste gefundene Parametersatz wird in jeder Kategorie als Vergleich aufgeführt. Dies führt zwar zu einer gewissen Redundanz in der Tabelle, erleichtert aber die Interpretation und den Vergleich mit der jeweiligen Parametermodifikation.

Zuerst sind die Ergebnisse der lokalen Ähnlichkeitsfunktionen jeweils für Farb- und Grauwertbilder angegeben. Die Fenstergröße bei den fensterbasierten Funktionen beträgt 5×5. Das Grauwertbild wurde aus den Intensitäten des Farbbildes berechnet. Interessanter Weise sind nicht alle Verfahren auf Farbbildern besser als auf Graubildern. Insbesondere die häufig verwendeten SSD- und NCC-Funktionen produzieren auf Grauwerten sogar geringfügig bessere Ergebnisse als auf Farbbildern. Gleiches gilt auch für das BT-Verfahren. Bei SAD, MNCC und MI hingegen weisen die Farbbilder sowohl beim Fehler als auch bei der Standardabweichung bessere Werte auf und sind den Grauwertbildern vorzuziehen.

Generell ist festzustellen, dass SAD und SSD selbst bei identischem Farbraum nicht an die Qualität der anderen Verfahren herankommen. Der Unterschied zwischen der modifizierten Kreuzkorrelation und der originalen Kreuzkorrelation ist zwar nicht groß, zeigt aber eine klare Tendenz zugunsten der modifizierten Variante. Beide Verfahren erreichen jedoch nicht die geringen Laufzeiten und hohen Abdeckungen der punktbasierten Verfahren. Das punktbasierte Verfahren von Birchfield und Tomasi [4] hat Vor- und Nachteile. Zwar ist es das schnellste Verfahren, allerdings ist der durchschnittliche Fehler im Vergleich zu den anderen Verfahre sehr hoch. Nur die SSD weist noch schlechtere Werte auf. Die Standardabweichung von BT befindet sich im Mittelfeld. Da dieses Verfahren allerdings ähnlich der SAD nicht helligkeits- und kontrastinvariant ist [26], wird es nicht zur Initialschätzung der MI verwendet.

Die Mutual Information wurde zweimal berechnet: Die Zeile 'MI' gibt die Resultate mit der vorgeschlagenen Initialschätzung mittels modifizierter Kreuzkorrelation an. Da in dieser synthetischen Analyse bekannt ist, dass die Bilder keinerlei Farbverfälschung haben, wurde die MI mit der identischen Funktion als Startwert zum Vergleich in der Zeile 'MI id'

berechnet. Die beiden Durchläufe sind nahezu identisch und die Laufzeitänderung von 1.2 Sekunden ist auf die Kreuzkorrelation in der kleinsten Auflösung zurückzuführen. In den folgenden Experimenten wird daher nur die MI als lokales Ähnlichkeitsmaß verwendet.

Bei der nächsten Reihe von Experimenten wird der Schwellwert für die Mindestähnlichkeit untersucht. Erwartet wurde, dass eine höhere Mindestähnlichkeit zwar zu geringeren Erkennungsraten führt, im Gegenzug jedoch auch der Fehler sinkt. Wie der Tabelle zu entnehmen ist, sinkt die Erkennungsrate bei steigender Mindestähnlichkeit wie erwartet, der Fehler aber steigt. Anscheinend sind die Informationen aus den Nachbarschaftsbedingungen so gut, dass sie auch nicht eindeutige Ähnlichkeitswerte mehr als ausgleicht. Daher wird auf eine Angabe der Mindestähnlichkeit in den folgendenden Experimenten verzichtet.

Die beiden Kostenfaktoren Q_1 und Q_2 sind nicht intuitiv zu erklären. Die Angabe erfolgt in Prozent von der maximalen Ähnlichkeit, das heißt, ein Wert von 10 entspricht einem Äquivalent von 0.1 der normierten Ähnlichkeitsfunktion. Insgesammt wurden 227 verschiedene Kombinationen von $Q_1 \epsilon \{1 \cdots 10\}$ und $Q_2 \epsilon \{2 \cdots 42\}$ untersucht. Auf eine vollständige Angabe aller Ergebnisse in Tabelle 12 wurde der Übersichtlichkeit halber verzichtet und nur für jedes Q_1die Kombinationen mit dem geringsten Fehler angegeben. Zeitlich liegen alle Werte auf demselben Nieveau. Der Abstand von Q_1 zu Q_2 liegt für alle Q_1 zwischen 18 und 26, wobei keine eindeutige Tendenz für steigende Werte von Q_1 auszumachen ist. Dies legt nahe, dass Tiefensprünge mit ca. 20% Ähnlichkeitsveränderung im Bild einhergehen. Der minimale Fehler befindet sich bei 10-36. Allerdings ist hier die Abdeckung etwas schlechter und auch die Standardabweichung geringfügig höher als bei der Kombination 8-30. Für die Experimente wurde daher der Wert 8-30 verwendet.

Die wichtigste Modifikation des SGMs liegt in der Verwendung von drei Bildern. Hierbei wird untersucht, wie sich Qualität und Performance des Dreibildfalls (3 B) im Vergleich zum Zweibildfall (2 B) verhalten. Des weiteren ist in Abschnitt 13.2.2 aufgeführt, dass die Konsistenzprüfung im Dreibildfall wahlweise für alle drei Bilder (3 K) oder nur für zwei Bilder (2 K) durchgeführt werden kann. Der Zweibildfall ist qualitativ signifikant schlechter und kann auch weniger Punkte zuordnen. Die Laufzeit im Zweibildfall ist um ca. 16 Sekunden bezieungsweise 23,5% geringer als im Dreibildfall. Allerdings ist gleichzeitig die zu untersuchende Datenmenge um 33% gefallen, was verdeutlicht, dass die vorgeschlagene Dreibildintegration sehr ressourcensparend ist und und gleichzeitig bessere Ergebnisse liefert. Der Unterschied zwischen den beiden Konsistenzprüfungen ist nicht so stark ausgeprägt. Der Geschwindigkeitsvorteil der 2 K-Messung ist darauf zurückzuführen, das der Konsistenzcheck häufig bereits bei der ersten Prüfung zum Erfolg führt und die zweite und dritte Prüfung nicht mehr erforderlich sind. Man muss je nach Anwendung abwägen, ob man zu Lasten der Genauigkeit mehr korrespondierene Punkte haben möchte oder nicht. Alle weiteren Experimente dieser Untersuchung werden mit minimal zwei konsistenten Punkten durchgeführt.

Das hierarchische Modell wird auf Qualität und Laufzeit geprüft. Die Bildkorrespondenzen werden mit unterschiedlichen Anfangsskalen von $1/8 = 2^4$ (L 4) bis $1 = 2^0$ (L 0) bezogen auf die Originalgröße berechnet. Es zeigt sich, dass die Anzahl der Skalenlevels nicht zu klein und nicht zu groß sein darf: Bei hoher Skalenanzahl ist das erste untersuchte Bild zu klein, um eine statistisch signifikante MI aufzubauen. Eventuell tritt auch die Fehlerfortpflanzung bei diesen großen Skalenleveln zu stark in Erscheinung, weshalb vermutlich in [26] von der Verwendung hierarchischer Modelle abgeraten wurde. Ist die Skalenanzahl gering (L 1) oder wird ganz auf ein hierarchisches Modell verzichtet (L 0), ist das Ergebnis ebenfalls nicht optimal. Das beste Ergebnis liegt bei L 2 mit einer Startauflösung von 320×256, da hier der richtige Kompromiss zwischen Informationsgehalt und Fehlerfortpflanzung gefunden wurde. Es zeigt sich, dass die Verwendung eines Hierarchischen Modells sehr wohl zur Verbesserung der Qualität beitragen kann, wenn das kleinste Bild nicht zu klein gewählt wurde. Die dramatische Laufzeitzunahme bei L 0 macht deutlich, wie stark das hierarchische Modell den Suchraum einschränkt. Die Laufzeit hat sich im Vergleich zu L 1 fast verdreifacht. Dieser Trend setzt sich noch von L 1 nach L 2 fort, wobei die Laufzeitverlängerung nur noch Faktor 1.29 beträgt. Bei den übrigen Skalen überwiegt der Programmoverhead, weshalb nur noch sehr geringe Laufzeitverbesserungen zu verzeichnen sind.

Die Verwendung von Gradienteninformation bringt eine geringfügige Verbesserung der Ergebnisse. Die Laufzeitverlängerung ist auf die Berechnung der Gradienten zurückzuführen.

Um den Einfluss der Medianreparatur und -glättung abzuschätzen, wurden verschiedene Fenstergrößen für die Medianreparatur (R) und die Medianglättung (S) untersucht. Mit zunehmender Maskengröße sinkt die Erkennungsrate und die Laufzeit steigt. Zwar sinkt auch der durchschnittliche Fehler marginal, allerdings bleibt die Standardabweichung auf sehr ähnlichem Nieveau. Wegen der Zunahme der Laufzeit sollte eine möglichst kleine Maskengröße gewählt werden, wobei die Größe für den Reparaturfilter größer als für den Glättungsfilter sein sollte.

Zuletzt wurden die Parameter für die Subpixelberechnung untersucht. Dazu wurde zunächst die Korrespondenzsuche ohne Subpixelberechnung durchgeführt. Danach wurden verschiedene Parameter zur Berechnung untersucht: die Berechnungsart der lokalen Subpixelkosten (SSD und NCC) und der Glattheitsparameter λ_{sub} aus Gleichung 119. Die angenommene Standardabweichung σ_{norm} wird mit 0.25 Punkten festgelegt. Es zeigt sich, dass die Verwendung von Subpixelinformation das Ergebnis signifikant verbessert. Allerdings benötigt die Subpixelberechnung ca. 24% mehr Rechenzeit, was aber auch daran liegt, dass die implementierte Subpixelberechnung noch nicht für Mehrkernsysteme optimiert wurde. Als lokale Kostenfunktion ist die SSD der Kreuzkorrelation vorzuziehen. Eine ausschließlich lokale Schätzung der Subpixelposition (λ_{sub}=0) führt zu unbeständigen Ergebnissen. Zwar sinkt der durchschnittliche Fehler im Vergleich zur Berechnung ohne Subpixel, allerdings steigt die Standardabweichung. Dies zeigt, wie stark die Subpixelberechnung selbst bei synthetischen Bildern mit Gauß´schem Rauschverhalten vom Rauschen abhängt. Die Einführung der Glättungsparameter verlängert zwar die Laufzeit, allerdings wird die Qualität erst dadurch eindeutig verbessert. Der durchschnittliche Fehler sinkt zwar mit zunehmender Glättung, die Standardabweichung jedoch steigt fast im selben Maße. Die einzelnen Werte für die Subpixelberechnung bei unterschiedlichen Glättungsparametern liegen so dicht beieinander, dass keine eindeutige Empfehlung gegeben werden kann, außer ihn größer als fünf zu wählen.

13.5.3 Middleburry Vision Benchmark

Um die vorgestellten Modifikationen mit anderen Korrespondenzsuchmethoden vergleichen zu können, werden sie mit dem bekannten Datensätzten der Middlebury Stereo Vision Gruppe verglichen [58]. Wegen der geringen Größe der Testbilder von maximal 450×375 Bildpunkten ist der hierarchische Ansatz hier nicht bedeutend und wurde abgeschaltet. Des weiteren sei angemerkt, dass in diesem Verfahren kein Verfahren zur Füllung der verdeckten Bereiche entwickelt wurde. Daher wurde bei der Analyse der Bilder die Bewertung auf "Nonocc" umgestellt, um die verdeckten Bereiche nicht auszuwerten. Die Ergebnisse sind im Screenshot der Abbildung 34 zu sehen. Die rote Markierung zeit die Ergebnisse dieser Arbeit, die blaue Markierung die Ergebnisse der ursprünglichen Implementierung des SGMs.

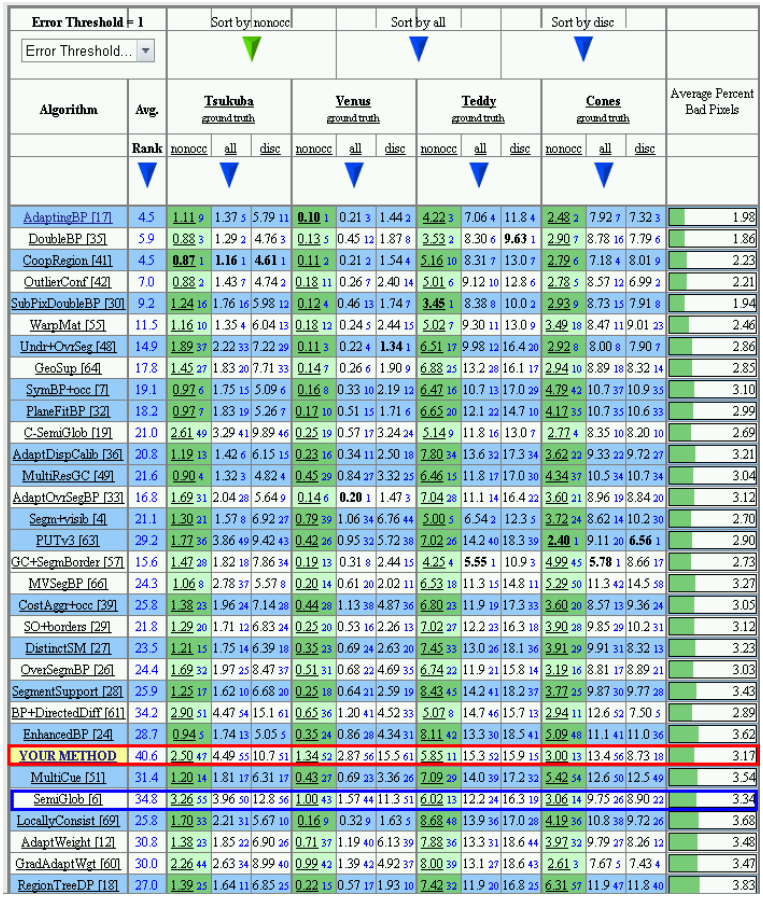
Obwohl die vorgestellte Technik primär für den Dreibildfall und sehr große Bilder optimiert wurde und gerade die MI durch die geringe Bildpunktanzahl der Testbilder benachteiligt ist, führen die vorgeschlagenen Verbesserungen zu besseren Ergebnissen als die Originalmethode. Hierbei sei angemerkt, dass das SGM in dieser Tabelle zwar nicht als eines der besten Verfahren abschneidet, es sich jedoch im Gegensatz zu vielen der vermeintlich besseren Verfahren als sehr vielseitig, robust und schnell erwiesen hat und für die vollautomatische Rekonstruktion besser geeignet ist. Die Rechenzeit dieser Implementierung für die einzelnen Bilder betrug zwischen vier und sieben Sekunden auf einem Athlon X2 mit $2 \times 2,4$GHz. Allerdings fallen bei so kurzen Laufzeiten der Verwaltungsaufwand und die Lese-/Schreibzugriffe des Algorithmus stark ins Gewicht, weshalb ein Vergleich mit Laufzeiten anderer Algorithmen nicht sehr aussagekräftig ist. Die Ergebnisse dieser Arbeit werden zur Veröffentlichung in die permanente Tabelle eingetragen.

Parameter	Erkennungs-rate	Fehler in Punkten	Std. Abw. in Punkten	Laufzeit in Sek
Lokale Ähnlichkeitsfunktion				
SAD - RGB	69.31	0.3331	0.3647	79.81
SAD - Grau	74.02	0.3585	0.3800	69.48
SSD - RGB	72.88	0.3646	0.3955	85.43
SSD - Grau	74.02	0.3453	0.3774	81.40
NCC - RGB	75.11	0.3487	0.3925	74.75
NCC- Grau	69.08	0.3414	0.3933	78.43
MNCC - RGB	74.84	0.3435	0.3903	75.19
MNCC - Grau	70.26	0.3438	0.4002	76.76
BT - RGB	77.18	0.3585	0.3865	65.16
BT - Grau	78.26	0.3336	0.3655	68.76
MI - RGB	83.83	0.3137	0.3396	68.16
MI - Grau	82.95	0.3174	0.3582	66.15
MI id - RGB	84.41	0.3111	0.3358	66.95
MI id - Grau	83.69	0.3126	0.3512	65.10
Farbe				
Grauwerte	82.95	0.3174	0.3582	66.15
Farbwerte	83.83	0.3137	0.3396	68.16
Mindestähnlichkeit				
0.0	83.83	0.3137	0.3396	68.16
0.2	83.72	0.3168	0.3432	68.13
0.4	83.95	0.3303	0.3512	68.36
0.6	83.23	0.3430	0.3544	68.80
0.8	67.60	0.3469	0.3725	84.34
Kostenfaktoren (Q_1 -Q_2)				
1 - 19	81.47	0.3419	0.3723	66.68
2 - 20	82.12	0.3359	0.3641	66.61
3 - 21	82.87	0.3274	0.3536	67.43
4 - 22	83.09	0.3236	0.3501	67.40
5 - 26	83.46	0.3181	0.3431	66.78
6 - 28	83.71	0.3144	0.3392	66.81
7 - 26	83.81	0.3148	0.3404	66.86
8 - 30	83.83	0.3137	0.3396	68.16
9 - 32	83.77	0.3138	0.3413	68.58
10 - 36	83.63	0.3135	0.3428	68.41

Tabelle 12: Quantitative Auswertung des SGM - Teil 1

Parameter	Erkennungs-rate	Fehler in Punkten	Std. Abw. in Punkten	Laufzeit in Sek
Zwei-/Dreibildfall und Konsistenzprüfung				
2 B	76.87	0.3554	0.4052	52.13
3 B / 2 K	83.83	0.3137	0.3396	68.16
3 B / 3 K	82.51	0.3001	0.3292	72.32
Hierarchisches Modell				
L 4	82.98	0.3080	0.3392	66.84
L 3	83.11	0.3075	0.3389	67.39
L 2	83.83	0.3137	0.3396	68.16
L 1	80.56	0.3225	0.3636	87.89
L 0	76.74	0.3377	0.3836	263.53
Gradienteninformation				
ohne	83.80	0.3138	0.3400	66.44
mit	83.83	0.3137	0.3396	68.16
Medianfilterung				
S=3, R=3	83.80	0.3141	0.3409	66.71
S=3, R=5	83.83	0.3137	0.3396	68.16
S=5, R=7	83.76	0.3125	0.3390	71.81
S=7, R=9	83.62	0.3129	0.3397	72.89
Subpixelberechnung				
Keine	82.17	0.4011	0.3509	54.81
SSD	83.80	0.3055	0.3333	66.81
XCOR	83.79	0.3133	0.3344	70.13
λ_{sub}=0	82.21	0.3656	0.3556	58.43
λ_{sub}=1	83.62	0.3153	0.3329	67.27
λ_{sub}=5	83.77	0.3094	0.3322	66.59
λ_{sub}=10	83.79	0.3072	0.3328	66.08
λ_{sub}=25	83.80	0.3055	0.3333	66.81
λ_{sub}=100	83.79	0.3039	0.3341	67.78

Tabelle 14: Quantitative Auswertung des SGM - Teil 2



Error Threshold = 1		Sort by nonocc				Sort by all				Sort by disc				
Error Threshold...														
Algorithm	**Avg.**	**Tsukuba** ground truth			**Venus** ground truth			**Teddy** ground truth			**Cones** ground truth			Average Percent Bad Pixels
	Rank	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	
AdaptingBP [17]	4.5	1.11 9	1.37 5	5.79 11	**0.10** 1	0.21 3	1.44 2	4.22 3	7.06 4	11.8 4	2.48 2	7.92 7	7.32 3	1.98
DoubleBP [35]	5.9	0.88 3	1.29 2	4.76 3	0.13 5	0.45 12	1.87 8	3.53 2	8.30 6	**9.63** 1	2.90 7	8.78 16	7.79 6	1.86
CoopRegion [41]	4.5	**0.87** 1	**1.16** 1	**4.61** 1	0.11 2	0.21 2	1.54 4	5.16 10	8.31 7	13.0 7	2.79 6	7.18 4	8.01 9	2.23
OutlierConf [42]	7.0	0.88 2	1.43 7	4.74 2	0.18 11	0.26 7	2.40 14	5.01 6	9.12 10	12.8 6	2.78 5	8.57 12	6.99 2	2.21
SubPixDoubleBP [30]	9.2	1.24 16	1.76 16	5.98 12	0.12 4	0.46 13	1.74 7	**3.45** 1	8.38 8	10.0 2	2.93 9	8.73 15	7.91 8	1.94
WarpMat [55]	11.5	1.16 10	1.35 4	6.04 13	0.18 12	0.24 5	2.44 15	5.02 7	9.30 11	13.0 9	3.49 18	8.47 11	9.01 23	2.46
Undr+OvrSeg [48]	14.9	1.89 37	2.22 33	7.22 29	0.11 3	0.22 4	**1.34** 1	6.51 17	9.98 12	16.4 20	2.92 8	8.00 8	7.90 7	2.86
GeoSup [64]	17.8	1.45 27	1.83 20	7.71 33	0.14 7	0.26 6	1.90 9	6.88 25	13.2 28	16.1 17	2.94 10	8.89 18	8.32 14	2.85
SymBP+occ [7]	19.1	0.97 6	1.75 15	5.09 6	0.16 8	0.33 10	2.19 12	6.47 16	10.7 13	17.0 29	4.79 42	10.7 37	10.9 35	3.10
PlaneFitBP [32]	18.2	0.97 7	1.83 19	5.26 7	0.17 10	0.51 15	1.71 6	6.65 20	12.1 22	14.7 10	4.17 35	10.7 35	10.6 33	2.99
C-SemiGlob [19]	21.0	2.61 49	3.29 41	9.89 46	0.25 19	0.57 17	3.24 24	5.14 9	11.8 16	13.0 7	2.77 4	8.35 10	8.20 10	2.69
AdaptDispCalib [36]	20.8	1.19 13	1.42 6	6.15 15	0.23 16	0.34 11	2.50 18	7.80 34	13.6 32	17.3 34	3.62 22	9.33 22	9.72 27	3.21
MultiResGC [49]	21.6	0.90 4	1.32 3	4.82 4	0.45 29	0.84 27	3.32 25	6.46 15	11.8 17	17.0 30	4.34 37	10.5 34	10.7 34	3.04
AdaptOvrSegBP [33]	16.8	1.69 31	2.04 28	5.64 9	0.14 6	**0.20** 1	1.47 3	7.04 28	11.1 14	16.4 22	3.60 21	8.96 19	8.84 20	3.12
Segm+visib [4]	21.1	1.30 21	1.57 8	6.92 27	0.79 39	1.06 34	6.76 44	5.00 5	6.54 2	12.3 5	3.72 24	8.62 14	10.2 30	2.70
PUTv3 [63]	29.2	1.77 36	3.86 49	9.42 43	0.42 26	0.95 32	5.72 38	7.02 26	14.2 40	18.3 39	**2.40** 1	9.11 20	**6.56** 1	2.90
GC+SegmBorder [57]	15.6	1.47 28	1.82 18	7.86 34	0.19 13	0.31 8	2.44 15	4.25 4	**5.55** 1	10.9 3	4.99 45	**5.78** 1	8.66 17	2.73
MVSegBP [66]	24.3	1.06 8	2.78 37	5.57 8	0.20 14	0.61 20	2.02 11	6.53 18	11.3 15	14.8 11	5.29 50	11.3 42	14.5 58	3.27
CostAggr+occ [39]	25.8	1.38 23	1.96 24	7.14 28	0.44 28	1.13 38	4.87 36	6.80 23	11.9 19	17.3 33	3.60 20	8.57 13	9.36 24	3.05
SO+borders [29]	21.8	1.29 20	1.71 12	6.83 24	0.25 20	0.53 16	2.26 13	7.02 27	12.2 23	16.3 18	3.90 28	9.85 29	10.2 31	3.12
DistinctSM [27]	23.5	1.21 15	1.75 14	6.39 18	0.35 23	0.69 24	2.63 20	7.45 33	13.0 26	18.1 36	3.91 29	9.91 31	8.32 13	3.23
OverSegmBP [26]	24.4	1.69 32	1.97 25	8.47 37	0.51 31	0.68 22	4.69 35	6.74 22	11.9 21	15.8 14	3.19 16	8.81 17	8.89 21	3.03
SegmentSupport [28]	25.9	1.25 17	1.62 10	6.68 20	0.25 18	0.64 21	2.59 19	8.43 45	14.2 41	18.2 37	3.77 25	9.87 30	9.77 28	3.43
BP+DirectedDiff [61]	34.2	2.90 51	4.47 54	15.1 61	0.65 36	1.20 41	4.52 33	5.07 8	14.7 46	15.7 13	2.94 11	12.6 52	7.50 5	2.89
EnhancedBP [24]	28.7	0.94 5	1.74 13	5.05 5	0.35 24	0.86 28	4.34 31	8.11 42	13.3 30	18.5 41	5.09 48	11.1 41	11.0 36	3.62
YOUR METHOD	40.6	2.50 47	4.49 55	10.7 51	1.34 52	2.87 56	15.5 61	5.85 11	15.3 52	15.9 15	3.00 13	13.4 56	8.73 18	3.17
MultiCue [51]	31.4	1.20 14	1.81 17	6.31 17	0.43 27	0.69 23	3.36 26	7.09 29	14.0 39	17.2 32	5.42 54	12.6 50	12.5 49	3.54
SemiGlob [6]	34.8	3.26 55	3.96 50	12.8 56	1.00 43	1.57 44	11.3 51	6.02 13	12.2 24	16.3 19	3.06 14	9.75 26	8.90 22	3.34
LocallyConsist [69]	25.8	1.70 33	2.21 31	5.67 10	0.16 9	0.32 9	1.63 5	8.68 48	13.9 36	17.0 28	4.19 36	10.8 38	9.72 26	3.68
AdaptWeight [12]	30.8	1.38 23	1.85 22	6.90 26	0.71 37	1.19 40	6.13 39	7.88 36	13.3 31	18.6 44	3.97 32	9.79 27	8.26 12	3.48
GradAdaptWgt [60]	30.0	2.26 44	2.63 34	8.99 40	0.99 42	1.39 42	4.92 37	8.00 39	13.1 27	18.6 43	2.61 3	7.67 5	7.43 4	3.47
RegionTreeDP [18]	27.0	1.39 25	1.64 11	6.85 25	0.22 15	0.57 17	1.93 10	7.42 32	11.9 20	16.8 25	6.31 57	11.9 47	11.8 40	3.83

Abbildung 34: Ergebnisse Middlebury Stereo

14 3D-Rekonstruktion

Die 3D-Rekonstruktion größerer Objekte aus verschiedenen Ansichten setzt voraus, dass die Szene unbeweglich ist. Eine einzelne Ansicht der Szene erlaubt nur eine 2,5D-Rekonstruktion, das heißt, verdeckte Bereiche werden nicht rekonstruiert und müssen durch Aufnahmen von verschiedenen anderen Standpunkten ergänzt werden. Ziel ist es daher, diese Standpunkte der Kameras vollautomatisch zu bestimmen, um anschließend die dichten 3D-Rekonstruktionen der Einzelansichten zusammenzuführen. Somit können auch größere Objekte und zunächst verdeckte Szenenbereiche erfasst werden. Für eine 3D-Rekonstruktion ist daher ein gemeinsames Koordinatensystem für alle Kameras nötig, wie es in Kapitel 11 bestimmt wurde. Das System verwendet als Referenz die erste Kamera des ersten Tripels und rekonstruiert die Szene bezüglich dieser Referenzkamera. Das Modell ist somit ausschließlich lokal referenziert. Der Maßstab der gesamten Rekonstruktion wird durch die Basislänge zwischen der ersten und der zweiten Kamera auf dem Rahmen festgelegt.

14.1 Triangulation

Die Triangulation der Raumpunkte erfolgt über die Lösung eines homogenen Gleichungssystems für die gegebenen Projektionsmatrizen P_i $(i\epsilon\,[1,N]\,,N\geq 2)$ mit den dazugehörigen Bildpunkten x_i:

$$
\begin{gathered}
x_i = P_i X \\
\Longleftrightarrow \\
P_i X - x_i = 0 \\
\Longleftrightarrow \\
\begin{bmatrix}
x_1^x P_1^{31} - P_1^{11} & x_1^x P_1^{32} - P_1^{12} & x_1^x P_1^{33} - P_1^{33} & x_1^x P_1^{41} - P_1^{41} \\
x_1^y P_1^{31} - P_1^{21} & x_1^y P_1^{32} - P_1^{22} & x_1^y P_1^{31} - P_1^{31} & x_1^y P_1^{41} - P_1^{41} \\
\vdots & \vdots & \vdots & \vdots \\
x_N^x P_N^{31} - P_N^{11} & x_N^x P_N^{32} - P_N^{12} & x_N^x P_N^{33} - P_N^{33} & x_N^x P_N^{41} - P^{41} \\
x_N^y P_N^{31} - P_N^{21} & x_N^y P_N^{32} - P_N^{22} & x_N^y P_N^{31} - P_N^{31} & x_N^y P_N^{41} - P_N^{41}
\end{bmatrix} \cdot \begin{bmatrix} X^x \\ X^y \\ X^z \\ X^w \end{bmatrix} = 0
\end{gathered}
\tag{121}
$$

Jeder Bildpunkt bringt demnach zwei neue Gleichungssysteme ein. Dies bedeutet für die drei unbekannten Raumkoordinaten, dass Gleichung 121 schon für zwei Kameras überbestimmt ist. Daher empfiehlt es sich, die Kleinste-Quadrate-Lösung von Gleichung 121 über eine SVD zu berechnen. Dieses Verfahren kann somit Korrespondenzen in beliebig vielen Kameras bearbeiten, wodurch die Fehler in Kameraorientierung und Korrspondenzlokalisation gleichmäßig verteilt werden. Allerdings kann schon ein Ausreißer das Gesamtergebnis erheblich verschlechtern, weil ein großer lokale Fehler in einer falschen Korrespondenz gleichmäßig auf alle korrekten Korrespondenzen verteilt wird. Ein überbestimmter Raumpunkt $\widetilde{X}$ erfüllt daher nur bei idealen Daten die Gleichung $x_i = P_i \tilde{X}$. Die Abweichung der Bildkoordinaten x_i von den Koordinaten des rückprojizierten Punktes $\widetilde{x_i}$ wird als Rückprojektionsfehler bezeichnet und berechnet sich aus:

$$
\begin{gathered}
\widetilde{x_i} = P_i \tilde{X} \\
err_{proj}\left(x_i, \tilde{x_i}\right) = \sqrt{\left(\widetilde{x_i^x} - x_i^x\right)^2 + \left(\widetilde{x_i^y} - x_i^y\right)^2}
\end{gathered}
\tag{122}
$$

Dieser Fehler ist ein gutes Maß für die Güte der Triangulierung. Es kann aber nicht unterschieden werden, ob der Rückprojektionsfehler aus einer fehlerhaften Korrespondenz, einer fehlerhaften Kameraorientierung oder einer fehlerhaften Triangulation entstanden ist. Wird bei einer Fehleranalyse nur qualitativ entschieden, welche Lösung die bessere ist, kann man auf den quadratischen Rückprojektionsfehler zurückgreifen, wodurch die rechenintensive Wurzelziehung entfällt.

14.1.1 Nachbarschaftsmodell aus den Bildern für Dreiecke

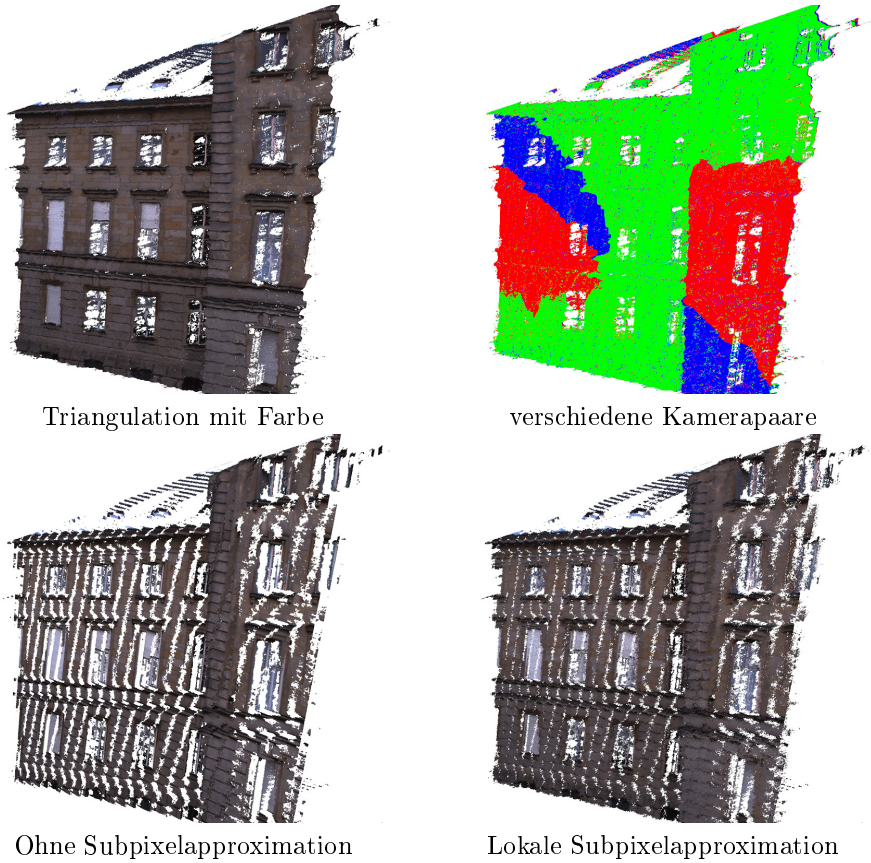
Da die 3D-Rekonstruktionen aus 2D-Bildern entstehen, kann man über die Bildnachbarschaft die 3D-Punkte zu Dreiecken vermaschen. Dabei wird geprüft, ob die benachbarten Bildpunkte korrespondierende Punkte haben, die wiederum benachbart sind, bei denen sich also die Disparitäten nur geringfügig unterscheiden. Mit diesen Informationen bilden nun drei oder vier benachbarte Punkte ein Polygon, so dass den Punkten nicht nur eine Fläche, sondern auch über Normale und Kamerarichtung eine eindeutige Orientierung zugeordnet werden kann. Zusätzlich kann das Referenzbild mit der triangulierten Rekonstruktion referenziert werden, um eine automatische Texturierung der Rekonstruktion zu erhalten.

14.2 Ergebnisse

Für die Auswertung der Triangulation wird ein Teilstück einer triangulierten Gebäudefassade untersucht. Die Berechnung der Korrespondenzen erfolgt über drei Bilder und das in Abschnitt 13.2 beschriebene Verfahren. Diese Korrespondenzen werden sowohl mit zwei als auch mit drei auf dem Rahmen montierten Kameras trianguliert. Um die Qualität des Subpixelverfahrens zu beurteilen, ist die Szene zum Vergleich auch ohne Subpixelberechnung sowie nur mit lokalen Subpixelkosten trianguliert worden. Die Ergebnisse sind in Abbildung 35 gezeigt. Der Standpunkt für die Abbildungsprojektion wurde so gewählt, dass er auf der Fassadenmitte einen rechten Winkel zu Aufnahmerichtung bildet.

Oben links ist die rekonstruierte Szene mit den aufgenommenen Farben gezeigt. Zu beobachten ist eine leichte Welligkeit auf der Oberfläche, die sich durch die Glättung im Subpixelbereich ergibt (vgl. unten links).

Oben rechts ist ein Ausschnitt aus den überlagerten Triangulationen der Kamerapaare 1/2 (rot), 1/3 (grün) und allen drei (blau) gezeigt. Die drei Farben der Tringulationen durchdringen sich gegenseitig und liegen geometrisch sehr nahe beieinander. Die Abstände der Punkte betragen maximal 4cm im Raum, was auf Disparitäten umgerechnet ca. 0.1 Pixel entspricht und damit unterhalb der angenommenen maximalen Messgenauigkeit von 0.2 Pixel liegt. Daher kann davon ausgegangen werden, dass die Orientierung der drei Kameras auf dem Rahmen inklusive der Skalierung mit dem hier beschriebenen Verfahren sehr genau ist.



Triangulation mit Farbe | verschiedene Kamerapaare

Ohne Subpixelapproximation | Lokale Subpixelapproximation

Abbildung 35: Triangulation mit mehreren Kameras und Subpixelapproximation

Der Vergleich des Referenzbildes oben links mit den unteren Bildern ohne Subpixelapproximation (links) und außschließlich mit lokaler Subpixelapproximation (rechts) zeigt, dass die Subpixelapproximation einen erheblichen Informationsgewinn bieten kann. Die ganzzahligen Disparitätssprünge sind deutlich als Streifen in der Fassade zu erkennen. Bei der lokalen Subpixelapproximation sind diese Streifen teilweise geschlossen. Allerdings kann häufig keine exakte lokale Subpixelinformation bestimmt werden, was durch die sichtbaren Stufen auch bei der ausschließlich lokalen Subpixelapproximation deutlich wird. Daher sollte die lokale Subpixelapproximation nicht ohne eine globale Glättung verwendet werden, um an Stellen ohne verwertbare Subpixelinformation zumindest die Glätte der Oberfläche zu gewährleisten.

15 Vollautomatische 3D-Rekonstruktion

Um eine vollautomatische Rekonstruktion durchzuführen, können die beschrieben Techniken zu einem Gesamtprogramm zusammengefügt werden. Als Voraussetzung müssen die radiale

Algorithmus 6 Vollautomatische 3D-Rekonstruktion

1. Erstelle radial entzerrte Bilder
   (a) Wandle die Bayer-Tiled-Bilder in Farbbilder um (Kapitel 4)
   (b) Generiere eine Transformationsmatrix pro Kamera für die radiale Entzerrung (Abschnitt 9.3)
   (c) Interpoliere alle Bildpositionen des entzerrten Bildes bikubisch (Gleichung 109)
2. Berechne die Förstnerpunkte und SIFT-Deskriptoren für jedes Bild (Abschnitt 10.1.1)
3. Bestimme Bildkorrespondenzen für einige Bildtripel (z.B. 10) (Abschnitt 10.1.2)
4. Berechne einen Trifokaltensor aus diesen Bildkorrespondenzen mit den in [22] beschriebenen Verfahren
5. Verfolge die Punktmerkmale über alle Triplets mit trifokalem Tracking (Kapitel 10)
6. Bestimme die Orientierung auf dem Rahmen aus einer zufälligen Auswahl der räumlichen Korrespondenzen des Trackings und den intrinsischen Kamerakalibrierungen (Abschnitt 11.2)
7. Skaliere die Kameras auf die gegebene Basislänge (analog zu Gleichung 74)
8. Extrahiere den Kamerapfad nach Kapitel 11
9. Führe einen Bündelblockausgleich dieser Kamerapositionen nach Abschnitt 11.5 durch
10. Für jedes $x-$te Bildtripel:
    (a) Rektifiziere das $x-$te Bildtripel mit Hilfe der Trackingkorrespondenzen (Kapitel 12)
    (b) Führe Semi-Global Matching auf dem rektifizierten Bildtripel durch (Kapitel 13)
    (c) Invertiere die Rektifizierung auf den gefundenen Korrespondenzen (Invertierung von Gleichung 93)
    (d) Trianguliere die Raumpunkte aus den SGM-Korrespondenzen mit den Kamerapositionen der Pfadschätzung (Kapitel 14)

Verzeichnung und die intrinsische Kalibrierung der drei Kameras bekannt sein. Wenn eine maßstabsgetreue Rekonstruktion erwünscht ist, wird der Abstand der Projektionszentren von zwei Kameras auf dem Rahmen für die Bestimmung der Basislänge benötigt.

15.1 Integration der einzelnen Module

Die Rekonstruktion beginnt bei synchronen Videorohdaten, die zunächst farbkonvertiert und radial entzerrt werden müssen. Dabei ist darauf zu achten, dass die Farbkonvertierer möglichst geringe systematische Artefakte produzieren, da sich diese bei der späteren Merkmalsextraktion störend auswirken. Die Genauigkeit des gesamten Systems hängt signifikant von der Genauigkeit der radialen Entzerrung ab, da ansonsten die Voraussetzung für die linientreue Abbildung nicht gegeben ist.

Ausgehend von drei radial entzerrten, synchronen Videoströmen und der intrinsischen Kalibrierung werden zunächst hochgenaue Punktmerkmale und eine universelle Beschreibung benötigt, um die initialen Korrespondenzen zwischen den Bildern zu finden. In Abschnitt 10.1.1 ist dafür eine Technik beschrieben, die mit ihren Standardparametern sehr genaue und gut auffindbare Merkmale für eine große Bandbreite an aufgenommenen Objekten liefert. Der Verzicht auf skalierungsinvariante Punktmerkmale ist für diese Anwendung

von Vorteil, da durch die hohe Bildfrequenz der Kameras ohnehin nur Bilder verglichen werden, die einander sehr ähnlich sind.

Die darauf folgende Zuordnung der Merkmale erfolgt über die Trifokalfiltertechnik. Zunächst werden die Punktmerkmale, wobei einige Ausreißer toleriert werden, mit der in Abschnitt 10.1.2 beschriebenen Technik zugeordnet. Alle benötigten Trifokaltensoren werden aus diesen Korrespondenzen vollautomatisch über GASAC bestimmt und die Außreißer in den Korrespondenzen gefiltert. Dabei liefern die guten Korrespondenzen automatisch die für die Guided-Matching-Technik des Trifokalfilters benötigten Grenzwerte für den erlaubten Sampsonabstand und den maximalen Rückprojektionsabstand.

Durch die Vielzahl an trilinearen Bedingungen, die diese Zuordnungen erfüllen müssen, sind die Trackingdaten praktisch frei von extremen Ausreißern. Der Restfehler bewegt sich im Rahmen der Rückprojektionsfehler, der erfahrungsgemäß vier Bildpunkte nicht überschreitet.

Aus diesen Trackingdaten kann die trifokale Rektifizierung bestimmt werden, um dichte Zuordnungen für die einzelnen Triplets mittels SGM zu bestimmen. Ferner dienen sie der Bestimmung der Kamerapositionierung auf dem Rahmen und der zeitlichen Kamerapositionen während der Aufnahme. Dabei ist besonders die Positionierung der Kameras auf dem Rahmen kritisch, da diese Rahmenpositionen zur Filterung der Ergebnisse der zeitlichen Kamerapositionsschätzung verwendet werden. Ein BBA optimiert abschließend die Positionsschätzung sämtlicher Kameras.

Zu guter Letzt werden die dichten Korrespondenzen mit den geschätzten Kamerapositionen trianguliert und die Raumpunkte in ein globales Koordinatensystem eingetragen.

Um die Datenmenge zu reduzieren, sollte nur eine bestimmte Menge von Kamerapositionen zur Triangulierung der 3D-Szene verwendet werden. Hierbei kann eine feste Intervalllänge x eingeführt werden. Je nach Geschwindigkeit der Kamera zum Objekt sollte x zwischen 1 und 100 liegen.

Dieser sequenzielle Ablauf der Rekonstruktionstechniken ist in Algorithmus 6 gezeigt.

15.2 Ergebnisse

Die zusammengefügten Einzelaufnahmen des Ernst-Reuter-Hauses ergeben eine Punktwolke von mehr als 5 Milliarden 3D-Punkten. Da diese Punktmenge mit der zur Verfügung stehenden Software nicht als Ganzes verarbeitet werden konnte, wurde eine Auswahl von 28 Millionen Punkten erzeugt. Teilansichten aus verschiedenen Perspektiven sind in Abbildung 36 zu sehen. Der Datensatz ist als ASCII-File im Datenformat XYZ-RGB in dem in Anhang B aufgeführten Link zur Verfügung gestellt. Der Maßstab im Datensatz beträgt einen Meter.

Die Rekonstruktion ist sehr detailreich und auch an den Übergangskanten von einer Ansicht zur nächsten sind keine Diskontinuitäten zu erkennen. Die Genauigkeit der Fassadenrekonstruktion wurde anhand des Grundrisses mit den gemessenen Abständen in den Punktwolken ermittelt. Das Raster der Fenster beträgt laut Grundriss exakt 3m. In der Rekonstruktion wurden Abstände von 2,97 - 3,02m gemessen, was einer Genauigkeit von 1% entspricht (vgl. Abbildung 37). Hierbei ist zu beachten, dass die Aufnahme der Fassade sehr oblique war (vgl. Abbildung 11), was auch die schrägen Sichtbereichskanten in Abbildung 36 oben bedingt. Daher sind die Messungen auf der Fassade nicht nur von der Position der $x-/y$-Ebene, sondern auch stark von der rekonstruierten Tiefe abhängig.

Um zu zeigen, dass auch komplexere Kamerabewegungen rekonstruierbar sind, wurde die Decke des ausgebrannten und mittlerweile eingestürzten Offizierskasinos in Rathenow mäanderförmig abgefilmt und rekonstruiert. Besonders schwierig bei dieser Aufnahme waren die Wendepunkte. An diesen Punkten musste der Rahmen gekippt werden, um auch die Ecken aufzunehmen, die durch Schutt unzugänglich waren. Die Decke hat eine Größe von 7.7×12.8 Metern. Der Abstand zwischen Decke und Kamera betrug 5-7 Meter. Die Sequenzlänge umfasste 3800 Bilder, wurde aber wegen der schlechten Beleuchtungsverhältnisse aus drei Aufnahmen mit unterschiedlich positionierten Scheinwerfern zusammengeschnitten, was einen Teil der Farbabweichungen erklärt. Der Kamerapfad und einige Ansichten sind in Abbildung 38, 39 und 40 zu sehen. Auch dieser Datensatz steht als Download in Anhang B zur Verfügung.

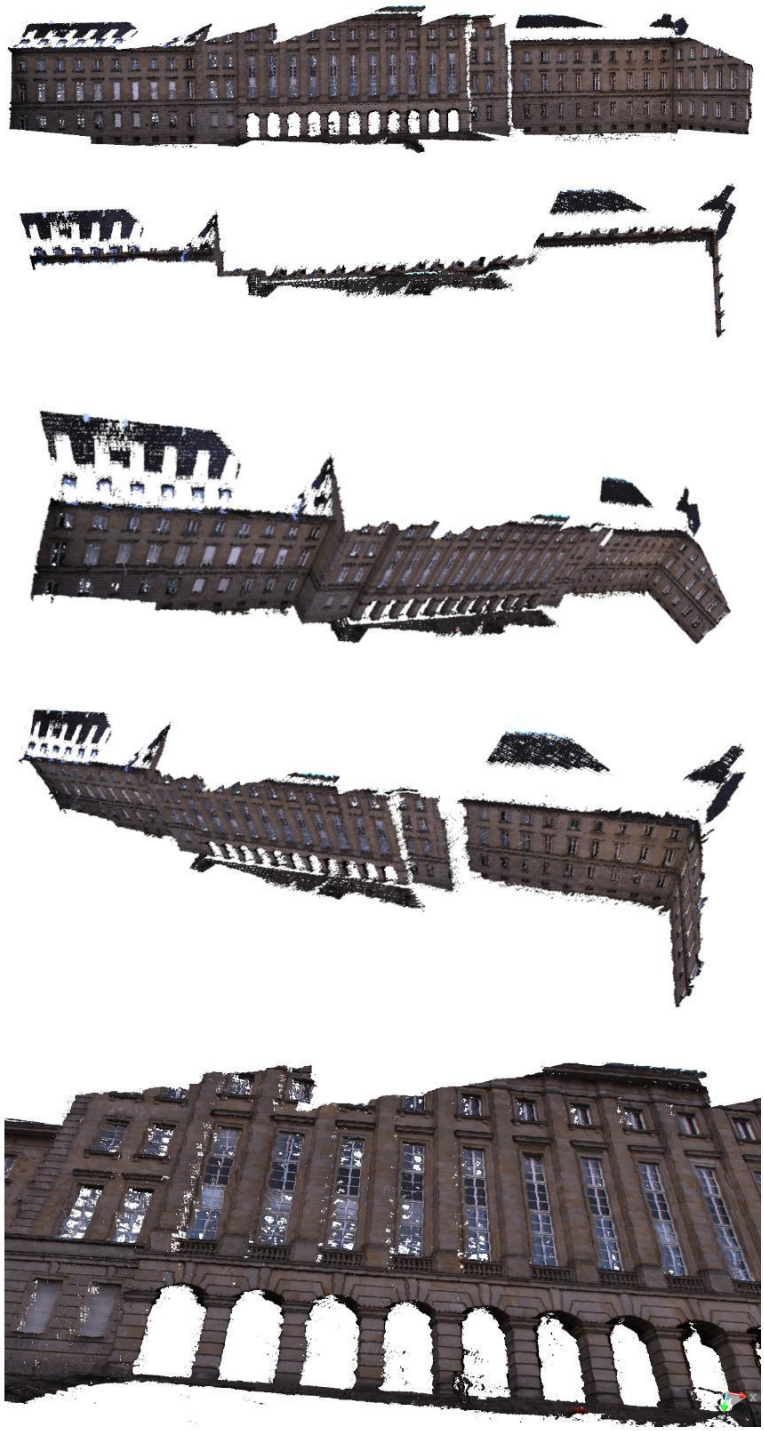
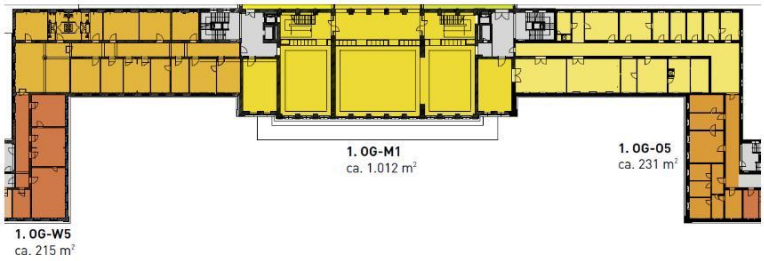
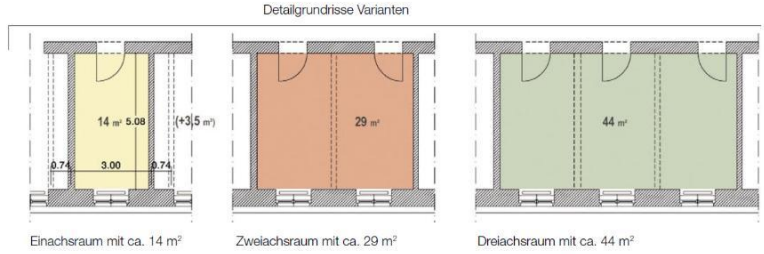


Abbildung 36: Rekonstruktion des Ernst-Reuter-Hauses

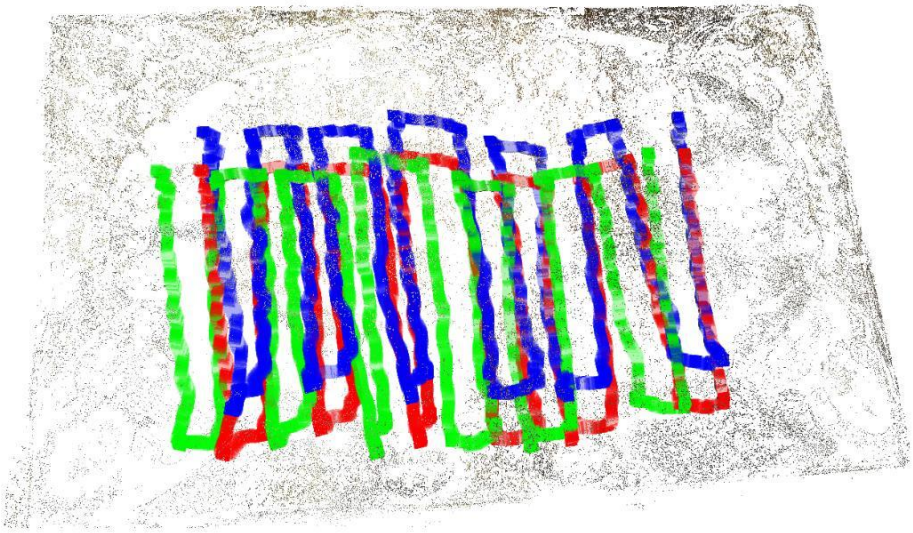




Grundriss

Messung

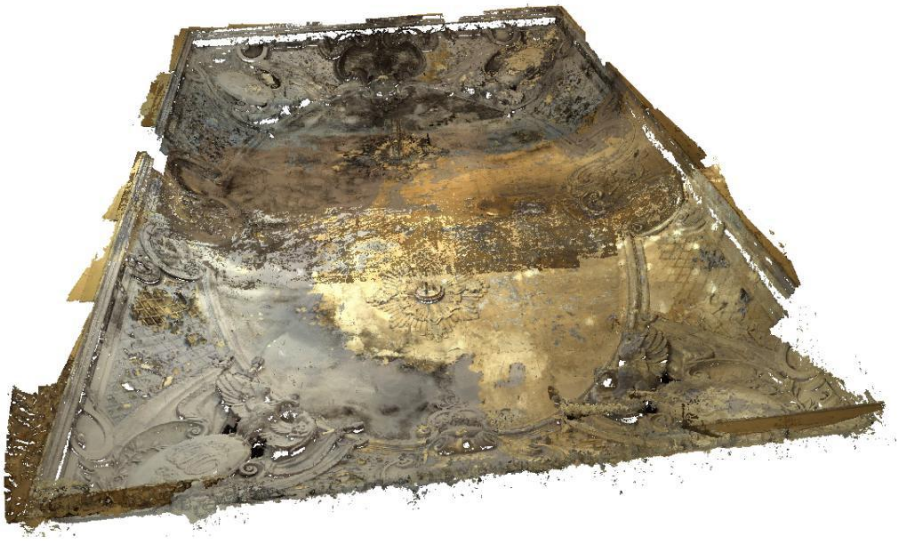
Abbildung 37: Messung der Rekonstruktionsgenauigkeit



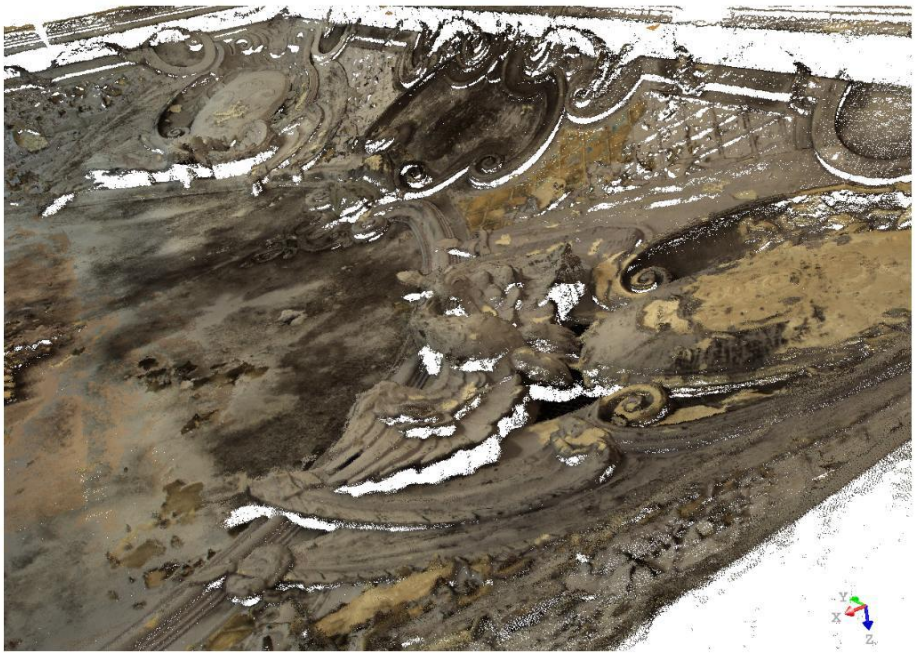
Kamerapfad mit Punktmerkmalen

Gesamtansicht der Decke

Abbildung 38: Pfad und Rekonstruktion des Offizierskasinos in Rathenow



Perspektivische Ansicht



Detailansicht Wappen

Abbildung 39: Ansichten der Rekonstruktion des Offizierskasino

Detailansicht Krone

Abbildung 40: Ansichten der Rekonstruktion des Offizierskasino II

Teil IV
Zusammenfassung und Ausblick

Es wurde gezeigt, dass das vorgestellte Gesamtprogramm sich für die vollautomatische Rekonstruktion von 3D-Szenen aus drei synchronen Videoströmen eignet. Die notwendigen Verfahren für Dreibildrektifizierung, effektive und dichte Stereokorrespondenzsuche, hochgenaue Merkmalslokalisation und -verfolgung, bedingte Kamerapfadschätzung und Bündelblockausgleich wurden vorgestellt und erfolgreich integriert.

16 Zusammenfassung

In diesem Kapitel werden die Ergebnisse zusammengefasst und kritisch diskutiert. Dabei werden die besonderen Punkte hervorgehoben, die das Gesamtprogramm vollautomatisch machen und gezeigt, wie die nötigen Parameter für die entsprechenden Teile automatisch bestimmt werden. Auf Schwächen oder kritische Fälle, die während der Entwicklung aufgetaucht sind, wird im Folgenden ebenfalls eingegangen.

16.1 Merkmalsextraktion

Alle vorgestellten Verfahren basieren zunächst auf markanten Punktmerkmalen, die im Falle der Rektifizierung als initiale Korrespondenzen benötgt werden oder im Falle der Pfadschätzung zunächst in den Videoströmen verfolgt werden müssen. Da der SIFT-Operator für höhere Skalierungen durch die Gaußglättungen ungenauer wird und in dieser Arbeit die Skalierungsinvarianz nicht benötigt wird, liefert die Kombination des Förstner-Operators mit dem SIFT-Dekriptor eine Möglichkeit der hochgenauen Merkmalslokalisierung mit guter Wiederfindbarkeit durch den SIFT-Deskriptor. Es wurde gezeigt, dass eine geeignete Subpixellokalisation Positionierungsgenauigkeiten von ca 0.25 Bildpunkten bei realen Daten erreichen kann. Die symmetrische Zuordnung der SIFT-Deskriptoren ist im Fall der Merkmalsverfolgung eine geeignete Erweiterung des ursprünglichen Zuordneverfahren von Lowe, um eine robustere Zuordnung mit großzügigen Verhältnisgrenzwerten zu erreichen.

Die in Abschnitt 10.1.1 vorgeschlagenen Parameter haben sich für eine vielzahl von Bildern als geeignet erwiesen. Da viele Parameter jedoch empirisch bestimmt wurden, wäre eine qualifizierte Herleitung dieser Parameter wünschenswert, um das theoretische Verständnis ihres Zusammenspiels vertiefen zu können. Des weiteren ist kein überzeugendes Verfahren für die Berechnung der SIFT-Deskriptoren aus Farbbildern bekannt, das der grauwertbasierten Technik überlegen ist.

16.2 Trifokales Tracking

Wegen der hohen Dimensionalität der SIFT-Deskriptoren können diese nicht mit Standardtechniken vorsortiert werden. Um dennoch die Zuordnungen bestimmen zu können, ist es nötig die Positionen der möglichen Kandidaten im Bildbereich so weit wie möglich einzuschränken. Dies erfolgt über die Verwendung des konstanten Trifokaltensors, der die Rahmengeometrie beschreibt. In der Regel können ca. 40% der im gemeinsamen Sichtbereich liegenden Punktmerkmale erfolgreich Merkmalen in den anderen Kameras zugeordnet werden, wobei der Rückprojektionsfehler dieser Punkttripel meist weniger als vier Bildpunkte beträgt. Darüber hinaus werden diese räumlich einheitlichen Bildkorrespondenzen für eine synchrone Kamerapfadschätzung benötigt, da nur so eine einheitliche Skalierung der drei Kamerapfade möglich ist. Durch die Berechnung der räumlichen und zeitlichen Trifokaltensoren ist es möglich, eine Analyse der Positionsabweichungen über robuste Filterung der Daten zu realisieren und die dafür benötigten Grenzwerte sogar automatisch anzupassen.

Allerdings ist die Berechnung der Trifokaltensoren sehr komplex und zeitintensiv. Das Verfolgen der Merkmale in 2100 Bildtripel dauert mehrere Stunden und ist daher noch nicht echtzeitfähig. Allerdings könnte eine Auslagerung diverser nummerischer Teilberechnungen

auf die vektororientieren GPUs moderner Grafikkarten hier entscheidende Impulse geben. Ein weiteres Problem ist das Wiederfinden von Merkmalen, die kurzfristig verdeckt wurden. Durch die Tifokaltensoren können Defekte nur über drei Bilder repariert werden. Treten die Defekte über längere Intervalle auf, kann die beschriebene Technik das verlorene Merkmal nicht wieder zuordnen und interpretiert es als neues Merkmal, wodurch der Pfad unterbrochen wird.

16.3 Kamerapfadschätzung

Das grundlegende Problem der Kamerapfadschätzung durch Elementarmatrixberechnung ist die Mehrdeutigkeit der Lösungen. Es kann nicht unterschieden werden, ob die Lösung durch ungeeignete Korrespondenzwahl falsch oder zu ungenau ist oder nur nicht richtig aus der Lösungsmenge ausgesucht wurde. Diese Problematik wird dadurch verschärft, dass die Basislänge zwischen dem Kamerapaar nicht bekannt ist und sogar Null sein kann. Durch die Bedingungen des Rahmens werden diese Probleme jedoch gelöst: Nur wenn die drei Kamerabewegungen nicht der Bewegungskoppelung auf dem Rahmen widersprechen, kann die Lösung korrekt sein. Darüber hinaus liefert der Rahmen durch die feste Montage der Kameras eine Skalierung, an die die Skalierung der Kamerabewegung für jedes Tripel angepasst werden kann. So wird eine Fehlerfortpflanzung der Skalierung, wie sie bei kamerapaarbasierten Pfadschätzungen auftreten kann, vermieden. Prinzipiell benötigt die Kamerapfadschätzung nur Toleranzwerte für die Bewegung der Kameras auf dem Rahmen. Die in Abschnitt 11.3.1 vorgeschlagenen Werte sind universell, so dass auch fünfmal schnellere Kamerabewegungen mit diesen Werten korrekt bestimmt wurden.

Allerdings erfolgt die Berechnung der Pfade im Grunde genommen unabhängig voneinander. Die Kalibirerinformationen werden erst im Nachhinein eingeführt, während die Bedingungen durch den Rahmen im projektiven Fall teilweise durch die Berechnung des Trifokaltensors schon während der Rechnung eingeführt wurden. Eine Lösung wäre ein "kalibrierter Trifokaltensor", der den Tensor nicht nur projektiv bestimmt. Eine explizite Berechnungstechnik für den kalibirerten Trifokaltensor ist bis zu dieser Veröffentlichung jedoch nicht bekannt. Daher ist es auch bei größtmöglicher Sorgfalt bei der Pfadberchnung notwendig, die Abweichungen der Kamerapositonen durch einen rechenintensiven Bündelblockausgleich zu minimieren.

16.4 Rektifizierung

Die Grundvorraussetzung für eine dichte Korrespondenzsuche ist eine Rektifizierung der Bilder, um eine symmetrische Zuordnung zu ermöglichen und Rasterisierungsartefakte durch die Gleikkommainterpolation entlang unterschiedlicher Epipolargeraden zu vermeiden. Die vorgestellte Rektifizierungstechnik für drei Bilder ermöglicht es, durch die symmetrischen Disparitäten ohne erheblichen Mehraufwand Korrespondenzen in drei Bildern gleichzeitig zu bestimmen. Ausgangspunkt dafür sind ausschließlich die initialen Korrespondenzen der gradentreuen Bilder und eine robuste Trifokaltensorbestimmung. Dabei wird eine größtmögliche Autonomie des Algorithmus gewährt, indem eine optimale Kamerapositionierung automatisch bestimmt wird und die robust bestimmten Tensoren so ausgewählt werden, dass die Bildverzerrung minimiert wird. Alle freien Parameter der Rektifizierung werden automatisch mit hilfe der vorgeschlagenen Techniken bestimmt. Als Ergebnis liegt neben den rektifizierten Bildern eine automatische Abschätzung des Disparitätenbereiches für die dichte Korrespondenzsuche vor.

Weichen die Kamerapositionen stark von den idealen Modelpositionen in Abbildung 24 ab, werden die rektifizierten Bilder immer stärker verzerrt. Gerade bei unterschiedlichen Basislängen der Kameras kann die dazu führen, dass durch die Verzerrung die Auflösung und damit der Informationsgehalt in den drei Bildern unterschiedlich ist, was später bei der Subpixelbestimmung Probleme hervorrufen kann. Dieser Effekt kann aber durch eine vernünftige Positionierung der Kameras auf dem Rahmen vermieden werden.

16.5 Semi-Global Matching

Die wichtigsten in dieser Arbeit vorgestellten Modifikationen des SGMs sind die Erweiterung der Korrespondenzsuche auf drei Bilder ohne Komplexitätssteigerung und das hierarchische Modell zur Verkleinerung des Suchraumes. Durch die Erweiterung auf den Dreibildfall kann die Robustkeit der Zuordnungen gesteigert werden, ohne die Parameter für jedes Bilderpaar einzeln zu optimieren. Durch die trinokulare Rektifizierung und die darin enthaltene Symmetrie der Disparitäten kann die Berechnung mit nur minimalem Mehraufwand durchgeführt werden. Das hierarchische Modell ermöglicht es große Bilder mit einem weiten Suchbereich zügig zu analysieren. Der tatsächlich betrachtete Suchbereich umfasst in der größten Auflösung in der Regel weniger als 40% des gesamten Suchraumes. Dieses hierarichische Modell führt im Bereich der Architektur, in dem häufig große zusamenhängenden Flächen auftreten, sogar zu einer Qualitätssteigerung, da durch die Bildpyramide Informationen über zusammenhängende Bildbereiche an die nächste Pyramidenstufe weitergeleitet werden. Durch die robusten Techniken ist der Einfluß der einzelnen Parameter nicht mehr so stark und es kann mit einem universellen Parametersatz eine Vielzahl von Bildern bearbeitet werden.

Auch wenn das SGM sehr gut parallelisierbar ist, besteht ein Flaschenhals im Speicherzugriff. Da der Algorithmus sehr große Speicherbereiche für die Berechnung der Kosten benötigt, kann das SGM große Bilder nur auf Rechnersystemen mit mehr als 16GB Speicher effektiv bearbeiten, da sonst sehr häufig Teile des Speichers auf die langsamen Festplatten ausgelagert werden. Ein weiterer unangenehmer Effekt ist die Welligkeit in der Subpixelapproximation. Die Ursache für diese Welligkeit ist nicht ausreichend geklärt und liegt eventuell im Approximationsverfahren selbst begründet.

16.6 Zusammenfügen

Verfügt man über die dichten Korrespondenzen aus dem SGM und die dazugehörigen Kamerapositionen, kann die Szene durch Triangulation dieser Punkte rekonstruiert werden. Die Ergebnisse in Abschnitt 15.1 zeigen die hohe Dichte und Genauigkeit der Daten. Hierbei ist zu beachten, dass ein Tripel ca. 3 Millionen Punktkorrespondenzen liefert, wovon jedoch nur einige Hunderttausend noch nicht im direkt vorherigen Tripel zu sehen waren. Um diese doppelten Raumpunkte zu vermeiden, müssen diese doppelten Einträge entweder gefiltert werden, oder durch geschickte Auswahl der Bilder weitestgehend vermieden werden. Bei den gezeigten Aufnahmen wurde auf eine möglichst konstante Kamerageschwindigkeit geachtet. So konnte ein festes Intervall bestimmt werden, bei dem der Überlappungsbereich sehr gering war. Bei schwankenden Kamerageschwindigkeiten kann über die Analyse der Trackinginformation bestimmt werden, wann ein Bildtripel genügend neue Bildinhalte aufweist, um eine nicht allzu lückenhafte Rekonstruktion bei gleichzeitig möglichst wenigen doppelten Raumpunkten zu gewährleisten.

17 Ausblick

In dieser Arbeit wurden Lösungsstrategien für viele Probleme vorgestellt. Natürlich gibt es viele Themen und Details, die nur ansatzweise oder gar nicht behandelt wurden. Dieses Kapitel behandelt eine Reihe von Ideen für zukünftige Arbeiten, ungelöste Probleme und angedachte Lösungsstrategien.

17.1 Detektion von Kreisschlüssen

Die Ausgleichsrechnung kann signifikant verbessert werden, wenn Kreisschlüsse des Kamerapfades als solche erkannt werden können. Kreisschlüsse liegen vor, wenn gefundene Punktmerkmale, die bereits nicht mehr verfolgt wurden, in späteren Bildern wiedergefunden und einheitlich indiziert werden können. Dies tritt sowohl bei zeitweiligen Verdeckungen als auch bei sehr langen Aufnahmewegen in Mäander- oder Kreisform auf. Es ist allerdings nicht praktikabel jede Kamerakombination auf Kreisschlüsse zu überprüfen, da dies einen zu hohen Aufwand bedeuten würde. Daher wird vorgeschlagen, ein mehrstufiges Verfahren

zu verwenden: Zuerst werden eine Pfadschätzung und eine Triangulation der Punktmerkmale nach den hier beschriebenen Verfahren und anschließend ein Bündelblockausgleich ohne Kreisschußdetektion durchgeführt. Das Ergebniss kommt der optimalen Lösung recht nahe, daher kann nun versucht werden, Kandidaten für einen Kreisschluss zu finden. Dazu wird für jede Kamera bestimmt, in welchem Bildausschnitt Punktmerkmale gefunden wurden und in welchem minimalen und maximalen Abstand sich die trianguliertenPunkte zur Kamera befinden. Alle Punkte, die sich in diesem Abstandsbereich vor der Kamera befinden und ausschließlich von anderen Kameras gesehen wurden, werden nun in das Bild projiziert. Dabei ist zu beachten, dass diese Punkte eventuell nicht exakt trianguliert wurden und die Position leicht fehlerbehaftet ist. Punkte werden zu Kandidaten für die Kreisschlussuntersuchung, wenn ihre Projektionen sich im Bereich der bekannten Merkmale befinden. Allerdings muss nun bestimmt werden, welches Bild zum Vergleich herangezugen werden soll. Dazu kann ein Histogramm über das Vorkommen der Kandidaten erstellt werden. Das Bild, in das die meisten Kandidaten projiziert wurden, wird als ähnlichstes Bild angesehen und die Merkmale dieses Bildes können über das Zuordnungsverfahren von Abschnitt 10.1.2 verglichen und gegebenenfalls neu indiziert werden.

17.2 Bewegungsschätzung durch Kalmanfilter

Bei einigen Aufnahmen tritt die Problematik auf, dass eine oder mehrere Kameras verdeckt werden. In diesem Fall ist der Kamerapfad unterbrochen und eine Referenzierung der Positionen in ein einheitliches Koordinatensystem ist nicht direkt möglich. Für diesen Fall bietet das Framework des Kalmanfilter [31], wie es in vielen SLAM-Anwendungen [9] verwendet wird, die Möglichkeit, diese Ausfälle durch Bewegungsmodelle abzuschätzen, um den Pfad zumindest grob konsistent zu halten. Diese Schätzung kann dazu führen, dass durch Detektion von Kreisschlüssen (Abschnitt 17.1) wieder echte Punkteferenzen in den Bildern und daduch ein einheitliches Koordinatensystem entstehen. Zusätzlich könnten diese Informationen bei der Auswahl der Elementarmatrix (Abschnitt 11.3.1) und einer zusätzlichen Validierung der Kamerabewegung helfen.

17.3 Räumliche Korrespondenzvalidierung durch Projektion in das nächste Bild

Die räumliche Korrespondenzsuche aus Abschnitt 8 ist leider nicht frei von Ausreißern. Daher wurde ein Verfahren zur stärkeren Vernetzung von räumlichem und zeitlichem Stereo angedacht, mit dem nicht nur die Ausreißer gefiltert, sondern auch die Genauigkeit verbessert werden sollte. Die Idee dazu kommt aus der bekannten Bewegung des Kamerasystems. Liegen die räumlichen Korrespondenzen eines Bildtripels und die Orientierung dieser Kameras vor (Abschnitt 11.2), können die Punkte trianguliert werden. Wurden einige Punktmerkmale ins nächste Tripel verfolgt und ist die Bewegung der Kamera bekannt, kann die triangulierte Wolke in das neue Kameratripel projiziert werden. Dadurch entstehen vermutete Korrespondenzen, die durch eine räumliche Korrespondenzsuche in diesem neuen Bildtripel validiert werden können. Wurde eine Korrespondenz durch mehrere Tripel bestätigt, ist die Wahrscheinlichkeit, dass es sich um einen Ausreißer handelt, stark verringert. Zusätzlich kann auch die Zahl der verfolgten Punktmerkmale durch dieses Verfahren gesteigert werden, wodurch der Bündelausgleich in merkmalsarmen Abschnitten verbessert werden kann.

Auch die Berechnung der Subpixel könnte durch dieses Verfahren verbessert werden, da die Subpixelapproximation über zwei Bilder mit noch größerer Basis erfolgen könnte. Ferner wäre ein ausgleichendes Überlagern der korrespondierenden Bildausschnitte denkbar, wodurch der Anteil des Bildrauschens reduziert würde.

17.4 Ausdünnen der Punktwolken

Da das vorgestellte Verfahren viele Punkte im Überlappungsbereich zeitlich aufeinander folgender Kameras mehrfach trianguliert, sollte ein Verfahren für die intelligente Auswahl der zu triangulierenden Punkte entwickelt werden. Angedacht wurde ein Algorithmus, der

durch Rückprojektion der Raumpunkte in ein vorangegangenes Kameratripel bestimmt, welche Punkte im gemeinsamen Überlappungsbereich liegen und daher nicht zur Punktwolke hinzugefügt werden sollten. Da für dieses Verfahren ähnliche Techniken wie die vorgeschlagene Erweiterung zur räumlichen Korrespondenzvalidierung benötigt werden, sollten diese zwei Themen nicht losgelöst voneinander betrachtet werden.

17.5 Auswertung von Farbbildern durch Mutual Information

Ein ungelöstes Problem besteht Interpretation von Farbinformation in der Kostenberechnung durch die Mutual Information (Abschnitt 13.1.2). In dieser Arbeit wird das arithmetische Mittel der drei Farbkanäle verwendet. Allerdings sind diese Kanäle nicht unabhängig voneinander, was durch das arithmetische Mittel nicht berücksichtigt wird. Ein vorstellbarer Lösungsweg wäre hier, einen anderen Farbraum zu wählen, z.B. den HSV-Farbraum, der Farbwert, Sättigung und Intensität beschreibt.

Danksagung

Diese Arbeit wurde von der Deutschen Forschungsgesellschaft finanziert. Ich bedanke mich für die in jeglicher Hinsicht guten Arbeitsbedingungen im Fachbereich Computer Vision & Remote Sensing an der Technischen Universität Berlin unter der Leitung von Prof. Dr. Olaf Hellwich. Mein besonderer Dank gilt Dr. Volker Rodehorst für die kompetente Betreuung dieser Arbeit und das zur Verfügung gestellte Handwerkszeug. Prof. Dr. Uwe Stilla danke ich audrücklich für die schnelle Begutachtung und die guten Nerven bei der abendeuerlichen Anreise zur wissenschaftlichen Aussprache. Ferner danke ich meinen Kollegen Holger Kumke und Ludwig Högner von der TU München, die uns bei der Erstellung der Testdaten tatkräftig unterstützt haben. Für inspirierende Gespräche und kritische Anmerkungen gilt mein besonderer Dank Prof. Dr. Wolfgang Förstner und Prof. Dr. Helmut Mayer. Nicht vergessen möchte ich meine Kollegen im „Sino-German Bundle Project“, mit denen ich interessante Diskussionen und nette Konferenzen in Deutschland und China erleben durfte, sowie meine Büronachbarn Saqib und Cornelius, die es jahrelang mit mir ausgehalten haben. Und last but not least geht mein Dank an Niki für das Lektorat sowie die moralische Unterstützung.

Anhang A: Namenskonvention und mathematische Grundlagen

Dieser Abschnitt definiert die in dieser Arbeit verwendeten Symbole und Notationen sowie die allgemeine Form bestimmter Transformationen, wie Scherung, Translation, Skalierung und Rotation. Für weitere Definitionen, z. B. Determinantenberechnung, Eigenschaften der Singulärwertzerlegung (*singlar value decomposition - svd*) oder der Eigensystemzerlegung wird auf weiterführende mathematische Werke wie [52, 65, 12] verwiesen.

Symbole und Notation

Name	Symbol	Erläuterung
Matrix	A	Kursive Großbuchstaben
Matrixelement	A^{xy}	Element der Matrix A in Zeile x und Spalte y
Vektor	a	Kursive Kleinbuchstaben
Vektorelement	a^x	Element x des Vektors
Variable	$thresh$	Kursive Kleinbuchstaben, Name ist im Text angegeben
Raumpunkte	X	Großes X, lokale Abweichung ist im Text angegeben
Bildpunkte	x	Kleines x, lokale Abweichung ist im Text angegeben
Koordinatenwert	x^xoder X^z	Koordinatenachse ist der obere Index
Skalarprodukt	$a \cdot b$	Skalarprodukt zweier Vektoren
Kreuzprodukt	$a \times b$	Kreutzprodukt zweier Vektoren
Schiefsymmetrisches Produkt	$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_\times$	Das schiefsymmetrische Produkt: $\begin{bmatrix} x \\ y \\ z \end{bmatrix}_\times = \begin{bmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{bmatrix}$
Vektorenlänge	$\|a\|$	Länge eines n-Vektors $\|a\| = \sqrt{(a^1)^2 + (a^2)^2 \cdots + (a^n)^2}$
Determinante	$det\,(A)$	Determinante von Matrix A
Spur	$spur(A)$	Summe der Diagonalelemente: $spur\,(A) = \sum_{i=1}^{n} A^{ii}$
Faltung	$f \otimes g$	Mathematische Faltung der Funktion f mit der Funktion g
Untermatrix	$reduce(P, x)$	Untermatrix von P ohne die x-te Spalte
Kameraindex (Raum, Zeit)	P_i^j	Räumlicher Index auf dem Rahmen: i, Zeitlicher Index (Tripel): j

2D-Homographien

Name	Form
Translation	$\begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}$
Rotation	$\begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix}$
Scherung in x-Richtung	$\begin{bmatrix} 1 & m_x & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
Scherung in y-Richtung	$\begin{bmatrix} 1 & 0 & 0 \\ m_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
Skalierung	$\begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$

3D-Homographien

Name	Form
Translation	$\begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix}$
Rotation um die x-Achse	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha & 0 \\ 0 & \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
Rotation um die y-Achse	$\begin{bmatrix} \cos\alpha & 0 & \sin\alpha & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\alpha & 0 & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
Rotation um die z-Achse	$\begin{bmatrix} \cos\alpha & -\sin\alpha & 0 & 0 \\ \sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
Skalierung	$\begin{bmatrix} s_x & 0 & 0 & 0 \\ 0 & s_y & 0 & 0 \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

Anhang B: Weblinks

- KLT feature tracker (Birchfield): `http://www.ces.clemson.edu/~stb/klt`
  (Zugriff am 07. Mai 2009)
- KLT feature tracker (OpenCV): `http://sourceforge.net/projects/opencv/`
  (Zugriff am 27. August 2009)
- Lowe SIFT: `http://www.cs.ubc.ca/~lowe/keypoints`
  (Zugriff am 11.09.2009, Binary Version 4)
- Sparse Bundle Adjustment: `http://www.ics.forth.gr/~lourakis/sba/`
  (Zugriff am 25.09.2009, Version 1.6)
- ICP Algorithmus: `http://www.mathworks.com/matlabcentral/fileexchange/16766-iterative-closest-point-method-c`
  (Zugriff am 25.09.2009, Version vom 24 Apr 2008)
- Datensatz Ernst-Reuter-Haus: `http://srv-43-200.bv.tu-berlin.de/~matzeh/3d/ERH.zip`
- Datensatz Rathenow: `http://srv-43-200.bv.tu-berlin.de/~matzeh/3d/OKR.zip`

Literatur

[1] Aristoteles and H. Flashar. *Aristoteles - Werke in deutscher Übersetzung: Problemata Physica*. Akademie Verlag Berlin, 4. edition, 1991.

[2] H. Bay, T. Tuytelaars, and L. J. V. Gool. Surf: Speeded up robust features. In *ECCV (1)*, pages 404–417, 2006.

[3] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.

[4] S. Birchfield and C. Tomasi. Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, 35(3):269–293, 1999.

[5] D. G. R. Bradski and A. Kaehler. *Learning opencv, 1st edition*. O'Reilly Media, Inc., 2008.

[6] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, November 1986.

[7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, Second Edition*. The MIT Press and McGraw-Hill Book Company, 2001.

[8] R. Deriche, Z. Zhang, Q.-T. Luong, and O. Faugeras. Robust recovery of the epipolar geometry for an uncalibrated stereo rig. In *ECCV '94: Proceedings of the third European conference on Computer vision (vol. 1)*, pages 567–576, Secaucus, NJ, USA, 1994. Springer-Verlag New York, Inc.

[9] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-whyte, and M. Csorba. A solution to the simultaneous localization and map building (slam) problem. *IEEE Transactions on Robotics and Automation*, 17:229–241, 2001.

[10] G. Egnal. Mutual information as a stereo correspondence measure. *Technical Report MS-CIS-00-20, University of Pennsylvania*, 2000.

[11] M. Felsberg and U. Köthe. Get: The connection between monogenic scale-space and gaussian derivatives. In R. Kimmel, N. Sochen, and J. Weickert, editors, *Scale Space and PDE Mathods in Computer Vision*, volume 3459 of *LNCS*, pages 192–203. Springer, 2005.

[12] G. Fischer. *Lineare Algebra*. Rowohlt-Taschenbuch-Verlag, Reinbek bei Hamburg, 1975.

[13] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[14] M. A. Föstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centers of circular features. In *ISPRS Intercommission Workshop*, Interlaken, Switzerland, 1987.

[15] S. Gehrig and U. Franke. Improving stereo sub-pixel accuracy for long range stereo. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–7, Oct. 2007.

[16] J. C. Glove. Manual of photogrammetry fifth edition. *American Society for Photogrammetry and Remote Sensing*, 2004.

[17] B. K. Gunturk, J. Glotzbach, Y. Altunbasak, R. W. Schafer, and R. M. Mersereau. Demosaicing: color filter array interpolation. In *IEEE Signal Processing Magazine*, pages 44–54, 2005.

[18] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[19] R. I. Hartley. In defense of the eight-point algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(6):580–593, 1997.

[20] R. I. Hartley. Chirality. *International Journal of Computer Vision*, 26(1):41–61, 1998.

[21] R. I. Hartley. Theory and practice of projective rectification. *International Journal of Computer Vision*, 35(2):115–127, 1999.

[22] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[23] M. Heinrichs, O. Hellwich, and V. Rodehorst. Efficient semi-global matching for trinocular stereo. In *Proc. of Photogrammetric Image Analysis (PIA 2007)*, volume 36, Part 3/W49A of *Int. Archives of Photogrammetry, Remote Sensing and Spatial Information Science*, pages 185–190, Munich, September 2007.

[24] M. Heinrichs, O. Hellwich, and V. Rodehorst. Robust spatio-temporal feature tracking. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.*, XXXVII. Part B3a.:51–56, 2008.

[25] M. Heinrichs and V. Rodehorst. Trinocular rectification for various camera setups. In *Symp. of ISPRS Commission III - Photogrammetric Computer Vision PCV'06*, pages 43–48, 2006.

[26] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 807–814, Washington, DC, USA, 2005. IEEE Computer Society.

[27] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):328–341, 2008.

[28] B. K. P. Horn. Recovering baseline and orientation from essential matrix. *Journal of the Optical Society of America*, 1990.

[29] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[30] S. Inglis, H. W. Thimbleby, and I. H. Witten. Displaying 3d images: Algorithms for single image random dot stereograms. *IEEE Computer*, 27:38–48, 1994.

[31] R. E. Kalman. A new approach to linear filtering and prediction problems.

[32] Y. Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. *Computer Vision and Pattern Recognition, 2004. Proceedings CVPR '04., 2004 IEEE Computer Society Conference on*, 2:506–613, Jun 2004.

[33] J. Kim, V. Kolmogorov, and R. Zabih. Visual correspondence using energy minimization and mutual information. In *International Conference on Computer Vision*, pages 1033–1040, 2003.

[34] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 15–18, Washington, DC, USA, 2006. IEEE Computer Society.

[35] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proceedings of IEEE International Conference on Computer Vision*, pages 508–515, 2001.

[36] C. Lei, J. Selzer, and Y.-H. Yang. Region-tree based stereo using dynamic programming optimization. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2378–2385, Washington, DC, USA, 2006. IEEE Computer Society.

[37] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, Sep. 1981.

[38] M. Lourakis and A. Argyros. The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm. Technical Report 340, Institute of Computer Science - FORTH, Heraklion, Crete, Greece, Aug. 2004.

[39] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[40] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI81*, pages 674–679, 1981.

[41] A. Lukin and D. Kubasov. An improved demosaicing algorithm. *International Conference on Computer Graphics and Vision GraphiCon*, 2004.

[42] F. Maes, A. Collignon, D. Vandeermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions in Medical Imaging*, 16:187–198, 1997.

[43] K. Mikolajczyk and J. Matas. Improving descriptors for fast tree matching by optimal linear projection. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[44] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. volume 27, pages 1615–1630, Los Alamitos, CA, USA, 2005. IEEE Computer Society.

[45] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(6):756–777, 2004.

[46] D. Nistér. Untwisting a projective reconstruction. *International Journal of Computer Vision*, 60(2):165–183, 2004.

[47] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. volume 1, pages I–652–I–659 Vol.1, 2004.

[48] D. Nistér and H. Stewénius. A minimal solution to the generalised 3-point pose problem. *J. Math. Imaging Vis.*, 27(1):67–79, 2007.

[49] D. Oram. Rectification for any epipolar geometry. In *British Machine Vision Conference*, 2001.

[50] J. Philip. Critical point configurations of the 5-, 6-, 7-, and 8-point algorithms for relative orientation. *Technical Report TRITA-MAT-1998-MA-13, Dept. of Mathematics, Royal Inst. of Tech. Stockholm*, 1998.

[51] J. Pluim, J. Maintz, and M. Viergever. Image registration by maximization of combined mutual information and gradient information. *Medical Imaging, IEEE Transactions on*, 19(8):809–814, Aug 2000.

[52] A. Quarteroni and F. Saleri. *Wissenschaftliches Rechnen mit MATLAB.* Springer-Verlag Berlin Heidelberg, 2006.

[53] V. Rodehorst. *Photogrammetrische 3D-Rekonstruktion im Nahbereich durch Auto-Kalibrierung mit projektiver Geometrie.* Ph.d. dissertation, Berlin University of Technology, June 2004.

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

[55] V. Rodehorst and O. Hellwich. Genetic algorithm sample consensus (gasac) - a parallel strategy for robust parameter estimation. In *Proc. of the Int. Workshop 25 Years of RANSAC held with IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2006)*, IEEE Computer Society, pages 1–8, New York, USA, June 2006.

[56] V. Rodehorst and A. Koschan. Comparison and evaluation of feature point detectors. *Proc. of 5th Turkish-German Joint Geodetic Days*, 2006.

[57] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *Proceedings of the Third Intl. Conf. on 3D Digital Imaging and Modeling*, pages 145–152, 2001.

[58] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2002.

[59] J. Shi and C. Tomasi. Good features to track. *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 593–600, Jun 1994.

[60] S. N. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc. Gpu-based video feature tracking and matching. Technical report, 2006.

[61] H. Stewénius, C. Engels, and D. Nistér. Recent developments on direct relative orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60:284–294, June 2006.

[62] P. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.

[63] H. von Sanden. *Die Bestimmung der Kernpunkte in der Photogrammetrien.* Dissertation an der Universität Göttingen, 1908.

[64] J. Weng, T. Huang, and N. Ahuja. Motion and structure from two perspective views: Algorithms, error analysis, and error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(5):451–476, 1989.

[65] D. Werner. *Funktionalanalysis.* Springer-Verlag Berlin Heidelberg, 2007.

[66] T. Werner. Constraint on five points in two images. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:203, 2003.

[67] H. Zhang, J. Čech, R. Šára, F. Wu, and Z. Hu. A linear trinocular rectification method for accurate stereoscopic matching. In R. Harvey and J. A. Bangham, editors, *BMVC 2003: Proceedings of the 14th British Machine Vision Conference*, volume 1, pages 281–290, London, UK, September 2003. British Machine Vision Association.

[68] Z. Zhang and T. Kanade. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27:161–195, 1998.

[69] K. Zuiderveld. *Contrast limited adaptive histogram equalization.* Academic Press Professional, Inc., San Diego, CA, USA, 1994.