Orientierung von Bildverbänden mit großer Basis

Jan Bartelsen

Institut für Angewandte Informatik

Universität der Bundeswehr München

17.12.2012

Universität der Bundeswehr München
Fakultät für Bauingenieurwesen und Umweltwissenschaften

Orientierung von Bildverbänden mit großer Basis

Dipl.-Inf. Jan Bartelsen

Vollständiger Abdruck der von der Fakultät für Bauingenieurwesen und Umweltwissenschaften der Universität der Bundeswehr zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Promotionsausschuss:

Vorsitzender: Univ.-Prof. Dr.-Ing. Friedrich S. Kröll
1. Berichterstatter: Univ.-Prof. Dr.-Ing. Helmut Mayer
2. Berichterstatter: Univ.-Prof. Dr.-Ing. Konrad Schindler (ETH Zürich)

Diese Dissertation wurde am 9.10.2012 bei der Universität der Bundeswehr München eingereicht.
Tag der mündlichen Prüfung: 17.12.2012

„Freiheit ist die Freiheit zu sagen, dass zwei plus zwei vier ist. Wenn das gewährt ist, folgt alles weitere.“

George Orwell – Nineteen Eighty-Four.

„Wir können nur eine kurze Distanz in die Zukunft blicken, aber dort können wir eine Menge sehen, was getan werden muss.“

Alan Mathison Turing – Computing Machinery and Intelligence (1950).

Danksagung

Ich bedanke mich bei Andreas Aßmuth für seine persönliche Unterstützung.

Der gute Englischunterricht von Victoria Baehren war mir eine große Hilfe.

Ohne die Mitarbeit von Patrick Reidelstürz wären viele Flugversuche nicht durchführbar gewesen.

Mein besonderer Dank gilt meinem früheren Kollegen Hai Huang, dessen Wissen über China und die Völker Asiens eine wichtige Bereicherung für mich war.

Kurzdarstellung

Die 3D Modellierung von urbanen Regionen und die Bestimmung von Höhen- und Landschaftsmodellen ist für Städteplanung, Tourismus, sowie Einsatzplanung von Spezialeinheiten des Militärs und der Polizei von zunehmender Bedeutung.

Digitale Bilder stellen aufgrund der niedrigen Kosten für Kameras und Speichermedien eine leicht verfügbare, massentaugliche Datenquelle dar. Der technische Aufwand zur Generierung der Bilder ist bei Bodenaufnahmen gering. Unter Verwendung von kleinen leichten unbemannten Flugsystemen (Unmanned Aircraft Systems – UAS) können auch Luftaufnahmen vergleichsweise einfach gewonnen werden. Koch *et al.* (1999); Pollefeys *et al.* (2000, 2004, 2008) haben gezeigt, dass unter ausschließlicher Verwendung digitaler Fotos und Methoden aus Photogrammetrie und Computer Vision eine 3D Rekonstruktion mit hohem Detaillierungsgrad möglich ist. Die Grundlage dafür bilden Verfahren zur Bildzuordnung, welche die relative Orientierung von Bildern ermöglichen.

In dieser Dissertation wird untersucht, wie die Robustheit der Bildzuordnung erhöht werden kann, um auch für Bildmengen mit großer Basis und ohne größere einschränkende Annahmen, die genaue Orientierung für eine detaillierte 3D Rekonstruktion bestimmen zu können. Dazu werden bestehende Methoden hinsichtlich ihrer Robustheit gegen verschiedene Störeinflüsse untersucht. Basierend auf den daraus resultierenden Erkenntnissen wird ein neues Verfahren zur robusten Bildzuordnung FASIAM – (Fast Accurate Scale Invariant Affine Matching) präsentiert und in ein Gesamtsystems zur 3D Rekonstruktion integriert. Zur Verifikation der Robustheit des Verfahrens werden Tests mit den Bilddatensätzen der Visual Geometry Group (2003) Oxford durchgeführt und mit Ergebnissen gängiger Verfahren verglichen. Das Gesamtsystem wird mittels größerer realistischer freihändig oder von kleinen UAS aus aufgenommenen Bildmengen getestet. Zudem wird untersucht, ob und mit welcher Genauigkeit absolute Orientierungen auf der Grundlage von verhältnismäßig ungenauen GPS-Messungen bestimmt werden können. Abschließend erfolgen eine Bewertung der Ergebnisse und ein Ausblick auf mögliche künftige Nutzungen und Weiterentwicklungen.

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Direkte 3D Rekonstruktion einer Szene aus Bildern

Aufgrund der Fortschritte in der Halbleitertechnologie sind digitale Kameras und Speichermedien stetig deutlich kostengünstiger und leistungsfähiger geworden. Digitale Fotos können mit geringem Aufwand erstellt werden und der Einsatz der Kameras ist an jedem Ort weltweit möglich. Folglich stellen digitale Fotos eine kostengünstige und massentaugliche Datenquelle dar. In Photogrammetrie und Computer Vision ist schon vor längerer Zeit gezeigt worden, dass unter ausschließlicher Verwendung digitaler Fotos dreidimensionale (3D) Rekonstruktion mit hohem Detaillierungsgrad möglich ist (Koch *et al.*, 1999; Pollefeys *et al.*, 2000, 2004, 2008).

Detaillierte 3D Modelle urbaner Regionen sind in zunehmenden Maß für die Stadtplanung, virtuellen Tourismus und digitale Filmproduktion von Bedeutung. Aber auch Spezialeinheiten der Polizei und des Militärs haben Bedarf an 3D Modellen von bebautem Gebiet. Die richtige Einschätzung von Sichtbereichen und Wirkungsmöglichkeiten von Waffensystemen ist auf der Grundlage von 2D Karten, Skizzen und Fotos sehr schwierig. 3D Modelle können diesen Teil der Einsatzplanung von Spezialeinheiten deutlich erleichtern. Zur Abbildung urbaner Regionen können von mobilen Erfassungssystemen, von luftgestützten Systemen aber auch von Hand aus aufgenommene Bilder verwendet werden. Es ist jedoch nicht immer möglich, aufwändige Aufnahmeverfahren mittels Luftbild-Befliegungen oder Befahrungen durch mobile Erfassungssysteme anzuwenden. So sind Überflüge von bewohnten Gebieten abhängig von den Bestimmungen eines jeweiligen Landes grundsätzlich genehmigungspflichtig, d.h. oft aufwändig und teuer und z.T. gar nicht erlaubt. Mobile Erfassungssysteme benötigen befahrbare Straßen. Auch der Einsatz aufwändiger und damit z.T. schwerer und unhandlicher optischer Systeme, wie Laserscanner, ist nicht uneingeschränkt

möglich. Deshalb kommt von Hand aus aufgenommenen digitalen Bildern z.T. eine große Bedeutung zu. Aufgrund der niedrigen Kosten und des geringen Aufwands sind sie für nahezu jede Region verfügbar.

Die im Internet frei verfügbare Information verzehnfacht sich nach gegenwärtigem Trend alle fünf Jahre. Neue Suchmethoden ermöglichen in zunehmendem Maße die Extraktion relevanter Information. Auch für Bildinformation wurden leistungsstarke Ansätze demonstriert (Li *et al.*, 2008; Nistér & Stewenius, 2006). Dementsprechend hoch ist das Potential durch große Mengen (frei) verfügbarer, digitaler Fotos und Videos. Automatische Verfahren zur 3D Rekonstruktion, welche lediglich handelsübliche PCs erfordern, ermöglichen unter Nutzung frei verfügbarer Datenquellen schnell und kostengünstig die Generierung von 3D Modellen urbaner Regionen. Gegenwärtig werden häufig manuelle und semi-automatische Methoden zur Modellierung von städtischen Gebieten verwendet. Abhängig vom Detaillierungsgrad kann der Zeitaufwand dafür sehr hoch sein. Je nach Vorgehensweise müssen für die Modellierung eines einzelnen Gebäudes bis zu drei Arbeitstage aufgebracht werden (Zadlo *et al.*, 2010).

Eine Kamera lässt sich, von Ausnahmen abgesehen, geometrisch durch eine Lochkamera modellieren. Bei dieser liegen während der Aufnahme jeder beobachtete Objektpunkt, das Projektionszentrum sowie der zugehörige Bildpunkt auf einer Geraden. Wurde ein Objektpunkt von unterschiedlichen Positionen aus aufgenommen, lässt sich bei bekannten Orientierungen der Kameras der Schnittpunkt der Geraden von den Bildpunkten über die Projektionszentren zum Objektpunkt und damit die Lokalisation des Objektpunktes relativ zu den Kameras bestimmen. Eine 3D-Rekonstruktion ist somit im Allgemeinen dann möglich, wenn die Bildpunkte eines Objektpunktes in Bildern verschiedener Aufnahmepunkte lokalisiert wurden.

Durch die Verwendung von Punkt-Operatoren und Bildzuordnungsverfahren ist es möglich, Punktkorrespondenzen in Bildpaaren automatisiert zu ermitteln. Fünf (Nistér, 2003) bzw. sieben (Hartley & Zisserman, 2004) Punktkorrespondenzen reichen aus, um im kalibrierten (Kamerakonstante / Brennweite und Hauptpunktlage sind bekannt) bzw. unkalibrierten Fall direkt, d.h. ohne Näherungen, die relative Orientierung zu bestimmen. Dies alles ermöglicht die direkte relative Orientierung der Bilder einer Szene ohne Verwendung von Näherungswerten, Markern oder Passpunkten. Mittels standardisierter Metainformation wie dem Exchangeable Image File Format (Exif) stehen in Bilddateien wertvolle Zusatzinformation wie Brennweite, Pixelgröße, Kameramodell oder GPS-Daten bereit.

Für diese Dissertation werden Arbeiten aus Photogrammetrie und Computer Vision einbezogen. Die Photogrammetrie ist Ende des 19. Jahrhunderts als Fachgebiet der Geodäsie entstanden

und beinhaltet Messmethoden und Auswerteverfahren, um aus Fotos und Messbildern eines Objektes seine räumliche Lage oder 3D Form zu bestimmen. Bei Computer Vision handelt es sich um einen wissenschaftlichen Bereich, welcher starke Beziehungen zur Informatik, Photogrammetrie, Signalverarbeitung und künstlicher Intelligenz aufweist. Forschungsaktivitäten zielen u.a. darauf ab, Objekte in Bildern zu detektieren und zu beschreiben, ihre Eigenschaften zu bestimmen und zu klassifizieren und auf Grundlage dieser Ergebnisse Entscheidungen zu treffen oder Prozesse zu steuern.

1.2 3D Modelle von urbanen Regionen

Abbildung 1.1: Gebäudemodelle der Stadt Stuttgart visualisiert mit Google Earth (Google, 2011). Im linken Bild geben die umliegenden Berge einen Eindruck des zugrunde liegenden digitalen Geländemodells.

Google Earth (Google, 2011) ist eine freie, über das Internet weltweit verfügbare Software zur Darstellung eines 3D Modells der Erde, welches von Google Incorporated herausgebracht wurde. Satelliten- und Luftbilder unterschiedlicher Auflösung werden mit Geodaten überlagert und auf einem digitalen Geländemodell aufgebracht. Die zugrunde liegenden Daten werden stetig aktualisiert und verbessert. Geometrische Abweichungen liegen für gewöhnlich im Bereich weniger Meter, sind jedoch, abhängig von der Qualität der Bilder und des Geländemodells, lokal sehr verschieden. Mit einem hohen Automatisierungsgrad wurden für komplette Stadtteile meist einfache Gebäudemodelle integriert (siehe Abb. 1.1), vereinzelt aber auch manuell erstellte, welche einen sehr hohen Detaillierungsgrad aufweisen. Die Modelle stehen über die Client- / Server Architektur dezentral zur Verfügung. Auf der Grundlage des offenen Standards Keyhole Markup Language

(KML) ist zudem eine lokale Integration von 3D Modellen möglich. Bing Maps (siehe Abb. 1.2)

Abbildung 1.2: Schlosspark der Stadt Karlsruhe visualisiert mit Microsoft Bing Maps (Microsoft, 2012). Es werden lediglich Luftbilder gezeigt, da die Option zur Darstellung von 3D Modellen Ende 2010 eingestellt wurde.

von der Firma Microsoft ist ebenfalls ein freies über das Internet verfügbares Programm mit einer Client- / Server Architektur. Ähnlich wie durch Google wurden dafür weltweit für größere Städte 3D Gebäudemodelle generiert, die Option zur Darstellung wurde jedoch Ende 2010 eingestellt (Microsoft, 2010).

Die zentrale Einrichtung des Geoinformationsdienstes der Bundeswehr ist das Amt für Geoinformationswesen (AGeoBw). Hochdetaillierte Gebäude- und Geländemodelle können die weltweiten Einsätze der Bundeswehr durch ressourcensparenden Einsatz, den Schutz von Leib und Leben von Soldaten und die Begrenzung von Kollateralschäden unterstützen. Zu diesem Zweck wurde das Verfahren AGeoBw 3D Welten zur Erstellung höchst auflösender True Ortho Mosaike und Oberflächenmodelle aus Sensordaten flugzeuggetragener Systeme und deren Visualisierung entwickelt (Winck, 2008). Es liefert ein hoch genaues 2,5D Modell einer Szene (siehe Abb. 1.3). Die Auswertung erfolgt mittels Hochleistungsrechensystemen und nimmt für Modelle typischer Größenordnung zwei Wochen Rechendauer in Anspruch. Für die Texturierung ist abhängig vom gewünschten Detaillierungsgrad ggf. erhebliche manuelle Nachbearbeitung erforderlich.

Gegenwärtig führen die Streitkräfte der NATO Simulationsumgebungen wie Virtual Battle Space 2 (siehe Abb. 1.4) ein. Diese sollen zur Unterstützung der Einsatzplanung und für die Ausbildung verwendet werden. Detaillierte 3D Gelände- und Gebäudemodelle sind hierbei eine Grundvoraussetzung für valide Simulationsergebnisse von Szenarien, die für niedrige Führungsebenen (Gruppen- bis Kompanieebene) ausgelegt sind. Automatisierte 3D Rekonstruktion von bebauten

Abbildung 1.3: Links und rechts oben: Übungsdorf Bonnland visualisiert in AGeoBw 3D Welten. Für dieses Resultat wurden ca. 4.300 Bilder einer digitalen Luftbildkamera (Vexcel Ultracam) verwendet. Die Auswertung erfolgte durch ein Hochleistungsrechensystem und nahm 14 Tage Rechendauer in Anspruch. Links unten: Bildüberdeckung für die Modellierung Bonnlands in 3D Welten. Rechts unten: Blick auf die Stadt Hammelburg. Mit freundlicher Unterstützung durch das AGeoBw Dezernat II 2 (2).

Gebieten kann genutzt werden, um für große Areale eine kostengünstige und aktuelle Modellierung zu erhalten. Auf der Grundlage von aufwändigen Kameras, Sensor- und Trägersystemen können gute Ergebnisse erzielt werden, die jedoch mit einem entsprechend hohem Kostenaufwand verbunden sind. Bilder von gewöhnlichen Kameras, die entweder von Hand oder von kleinen, leichten und billigen Trägersystemen aus aufgenommen werden, stehen in deutlich höherem Maße zur Verfügung. Unter Verwendung von Verfahren zur relativen Orientierung, die nicht auf zusätzliche Information von hoch genauen und damit sehr teuren Inertialen Navigationssystemen (INS)

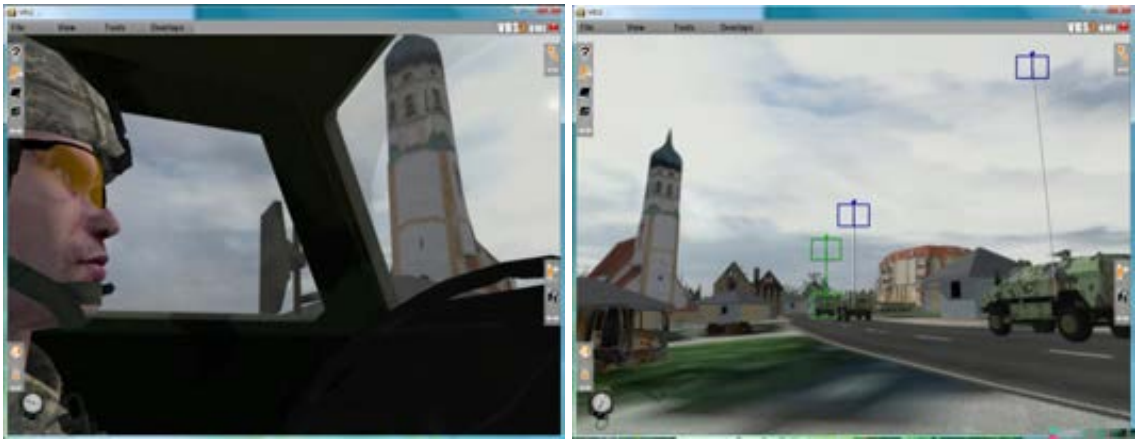


Abbildung 1.4: Simulationsumgebung Virtual Battle Space 2 – Fa. Bohemia Interactive Simulations.

angewiesen sind, könnte auf diese Weise eine detaillierte 3D Modellierung von bebautem Gebiet mit geringem Ressourcenaufwand erfolgen. Die Randbedingungen, die gängige Verfahren an Bildmengen stellen, sind jedoch nicht praxistauglich. Häufig werden hohe Bilddichten vorausgesetzt und oftmals führen bereits geringfügige Unterschiede in Bildpaaren dazu, dass eine Orientierung nicht mehr möglich ist. Auch wenn die entsprechenden Bilddichten vorliegen, liegt die Genauigkeit des durchschnittlichen Rückprojektionsfehlers in der Regel deutlich über einem Pixel, so dass die Verwendung der Orientierung für Verfahren zur Tiefenschätzung nur begrenzt oder gar nicht möglich ist.

Gegenstand dieser Dissertation ist es, für freihändig aufgenommene Bildsequenzen mit großer Basis, auch unter Verwendung verschiedener Kameras, eine zuverlässige, genaue und schnelle relative und z.T. auch absolute Orientierung zu ermöglichen, deren durchschnittlicher Rückprojektionsfehler deutlich unter einem Pixel liegt. Bis einschließlich Kapitel 3 werden bestehende Arbeiten erläutert und analysiert. Die folgenden Kapitel stellen die originären Anteile dieser Arbeit dar.

Zunächst werden Grundlagen eingeführt. Anschließend wird ausführlich betrachtet, welche Möglichkeiten und Grenzen bestehende Ansätze zur Bildzuordnung und 3D Rekonstruktion aufweisen. Basierend auf den daraus hervorgehenden Erkenntnissen wird ein neuartiger robuster Ansatz vorgestellt. Dieser wird anhand von quasi Benchmarktests und verschiedener Experimente evaluiert. Die Ergebnisse werden bewertet und abschließend folgen eine Zusammenfassung und ein Ausblick.

Kapitel 2

Grundlagen

In diesem Kapitel werden grundlegende Fakten erläutert, die zum Verständnis dieser Dissertation erforderlich sind. Dazu wird auf die mathematischen Grundlagen des Lochkameramodells und der Epipolargeometrie zur Rekonstruktion einer Szene, die 3D Ähnlichkeitstransformation, robuste Parameterschätzung und Bündelausgleichung sowie die Nutzung der Gaußfunktion zur Glättung der Bildfunktion eingegangen.

2.1 3D Rekonstruktion aus Bildern

Bei Bildern, welche über eine perspektive Abbildung generiert wurden, liegt die Information über räumliche Objekte in Form von 2D Daten vor. Im Weiteren wird angenommen, dass die verwendeten Bildmengen aus Ansichten von verschiedenen Kamerastandpunkten aus auf dasselbe starre Objekt bestehen. Auf dieser Grundlage ist eine 3D Rekonstruktion der Szene möglich (Hartley & Zisserman, 2004; Koch *et al.*, 1999; Pollefeys *et al.*, 2000).

Homogene Koordinaten
Homogene Koordinaten sind eine Erweiterung von euklidischen Koordinaten um eine weitere Koordinate und freie Skalierung.
Im Zweidimensionalen bedeutet dies:

$$\mathbf{x} = \lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} \text{ mit } x = \frac{u}{w},\ y = \frac{v}{w},\ \lambda \neq 0.$$

Und im Dreidimensionalen:

$$\mathbf{X} = \lambda \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} U \\ V \\ W \\ T \end{bmatrix} \text{ mit } X = \frac{U}{T},\ Y = \frac{V}{T},\ Z = \frac{W}{T},\ \lambda \neq 0.$$

Für $w = 0$ bzw. $T = 0$ werden x und y bzw. X, Y und Z beliebig groß und der Punkt liegt unendlich weit weg, seine Richtung kann aber dennoch angegeben werden.

Innere und äußere Orientierung

Eine Kamera wird durch zwei Arten von Parametern beschrieben: Die Parameter zur Beschreibung der Position des Projektionszentrums einer Kamera relativ zur Bildebene sowie deren affiner Verzerrung werden als Parameter der inneren Orientierung bezeichnet. Die Parameter zur Beschreibung der Position des Projektionszentrums und der Aufnahmerichtung relativ zum Aufnahmeobjekt nennt man Parameter der äußeren Orientierung (siehe Abb. 2.1). Zur inneren Orientierung, welche für gewöhnlich mittels einer 3×3 Kalibriermatrix (obere Dreiecksmatrix mit fünf Parametern) repräsentiert wird, zählen die Hauptpunktkoordinaten, Kamerakonstante, Skalierungsunterschied und Scherungsparameter. Typischerweise werden auch die Verzeichnungsparameter (siehe unten) der inneren Orientierung zugeordnet. Ist die innere Orientierung der Kamera bekannt, dann wird sie als **kalibriert** bezeichnet, ansonsten als **unkalibriert**. Die äußere Orientierung kann für zwei und mehr Bilder in relative und absolute Orientierung aufgeteilt werden.

Projektionsmatrix

Die 3×4 Projektionsmatrix $\mathbf{P}$ beschreibt die perspektive Abbildung eines 3D Objektpunktes in die Bildebene. Die Abbildung besteht aus Kalibrierung $\mathbf{K}$, Rotation $\boldsymbol{R}$, der Einheitsmatrix $\boldsymbol{I}$ und Translation t:

$$\mathbf{P} = \mathbf{K}\boldsymbol{R}[\boldsymbol{I}|-t].$$

Aufgrund der Homogenität hat die Projektionsmatrix nur 11 Freiheitsgrade. Wegen

$$\mathbf{K} = \begin{pmatrix} c & s & x_0 \\ 0 & c(1+m) & y_0 \\ 0 & 0 & 1 \end{pmatrix}$$

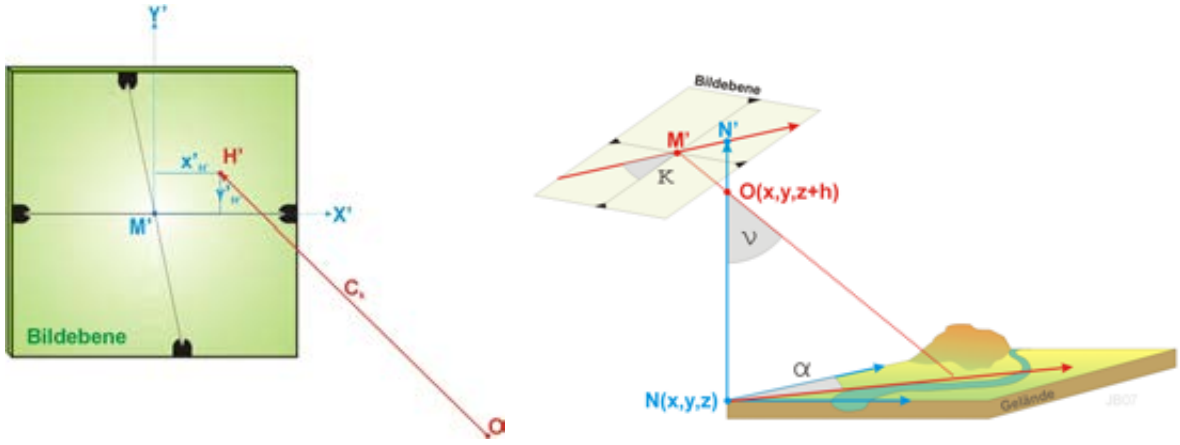


Abbildung 2.1: Dargestellt ist für die innere Orientierung (links) der Hauptpunkt H', welcher in der Regel nicht dem Mittelpunkt M' des Bildes entspricht. Der senkrechte Abstand zwischen dem Hauptpunkt und dem Projektionszentrum O wird als Kamerakonstante c_k bezeichnet. Für die äußere Orientierung (rechts) wird das Projektionszentrum O im Objektkoordinatensystem $(x, y, z+h)$ lokalisiert. Das Projektionszentrum befindet sich senkrecht über dem Nadirpunkt N. Die Rotation wird z.B. durch die drei voneinander unabhängige Drehwinkel Azimut α, Neigung ν und Kantung κ eindeutig beschrieben. Skizzen: Wikipedia – Die freie Enzyklopädie.

weist die Kalibriermatrix $\mathbf{K}$ mit der Kamerakonstante c, Skalierungsunterschied m, Scherung s und den Hauptpunktkoordinaten x_0, y_0 fünf, die Rotationsmatrix $\boldsymbol{R}$ sowie die Translation t weisen je drei Freiheitsgrade auf.

Radialverzeichnung

Die Verzeichnung ist ein geometrischer Abbildungsfehler optischer Systeme, welcher eine lokale Veränderung des Abbildungsmaßstabes verursacht. Wenn sich die Maßstabsänderung, wie häufig, in einer Änderung der Vergrößerung mit größer werdendem Abstand des Bildpunktes von der optischen Achse äußert, dann wird die Verzeichnung als radial bezeichnet. Wenn die Vergrößerung zu den Rändern der Bildebene zunimmt, findet man kissenförmige, im umgekehrten Fall tonnenförmige Verzeichnung vor. Neben radialer und der dazu senkrechten tangentialen Verzeichnung sind auch Verzeichnungen höherer Ordnung möglich. Für die Rekonstruktion der Aufnahmekonfiguration auf der Grundlage von Punktkorrespondenzen (siehe Abschnitt 3.2) ist eine Korrektur der Bildkoordinaten bezüglich der Verzeichnung erforderlich. Zur Korrektur des durch die Radialver-

zeichnung bedingten Fehlers kann beispielsweise folgende Näherung verwendet werden:

$$ds = k_2 * (r^2 - r_0^2) + k_4 * (r^4 - r_0^4).$$

Die zugehörigen Parameter k_2, k_4, r und r_0 werden im Weiteren als Verzeichnungsparameter bezeichnet. Hierbei ist r der Abstand vom Bildhauptpunkt und r_0 eine Konstante, die oft auf $\frac{2}{3}$ der Bildbreite gesetzt wird.

Relative und absolute Orientierung

Die relative Orientierung bestimmt die Translation und Rotation von Kameras im Raum zueinander. Auf dieser Grundlage kann für die Szene ein dreidimensionales, euklidisches Koordinatensystem bestimmt werden. In der Praxis lassen sich so zahlreiche Bilder zu einem Modell zusammenfügen (siehe Abb. 2.2). Da bei einer perspektiven Abbildung ohne Hinzunahme zusätzlicher Information der absolute Abstand zwischen Projektionszentrum und Objektpunkten nicht bestimmbar ist, ist der Maßstab des relativen Modells relativ zum Weltkoordinatensystem unbekannt. Das Modell aus der relativen Orientierung besitzt eine euklidische Geometrie, die räumliche Position und Orientierung im Weltkoordinatensystem ist wie der Maßstab nicht bekannt. Zur Bestimmung der verbliebenen Unbekannten der 3D Ähnlichkeitstransformation (siehe Abschnitt 2.2) werden typischerweise Passpunkte oder Positions- und Rotationsinformation für die Aufnahmen unter Verwendung des Global Positioning Systems (GPS) oder von Inertialen Navigationssyste-

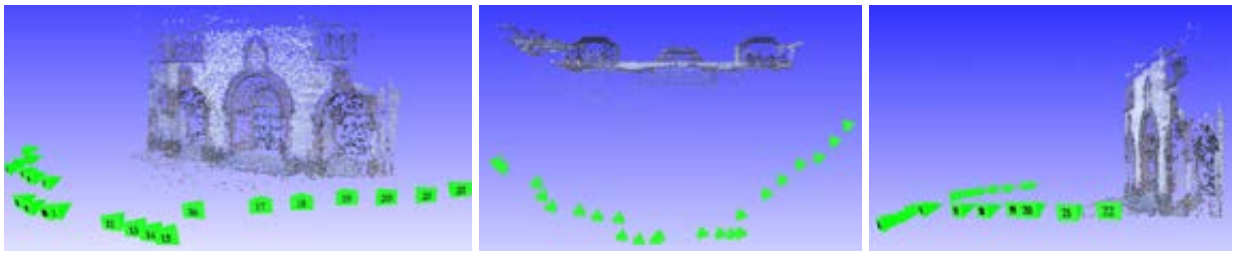


Abbildung 2.2: Das Ergebnis der relativen Orientierung von 23 Bildern (Herz-Jesu-R23, (Strecha *et al.*, 2008)), zeigt die Fassade der Kirche Herz-Jesu (Ettlingen) in Form einer 3D Punktwolke und die Kamerapositionen (grüne Pyramiden).

men (INS) verwendet.

Direkte Rekonstruktion einer Szene

Durch die Verwendung von Punkt-Operatoren und Zuordnungsmethoden (Förstner & Gülch, 1987; Harris & Stephens, 1988; Lowe, 2004) ist es möglich, in Bildpaaren korrespondierende Punkte automatisiert zu ermitteln (siehe Kapitel 3). Die geometrische Beziehung zwischen zwei Kamerabildern kann mittels der Epipolargeometrie (siehe Abb. 2.3) repräsentiert werden. Fünf (Nistér, 2003) bzw. sieben (Hartley & Zisserman, 2004) Punktkorrespondenzen reichen aus, um im kalibrierten bzw. unkalibrierten Fall direkt, d.h. ohne Näherungen, die Orientierung bzw. die Epipolarlinien zu bestimmen. Zusammen mit der automatisierten Bestimmung der Punktkorrespondenzen ist damit die direkte relative Orientierung einer Szene ohne Verwendung von Näherungswerten, Markern oder Passpunkten möglich. Die Epipolargeometrie wird beschrieben durch die (homogene) Fun-

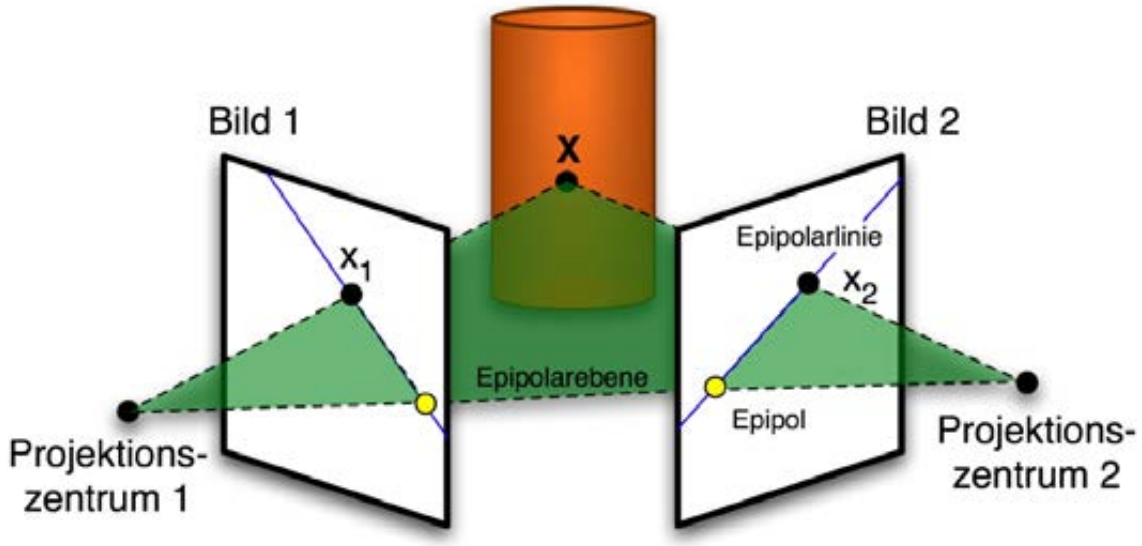


Abbildung 2.3: Epipolargeometrie einer Szene. Der Objektpunkt X ist sowohl im Bild 1 als auch im Bild 2 sichtbar. Zur einfacheren Darstellung wurden die Bildebenen jeweils vor die Projektionszentren eingezeichnet. Bei bekannter Epipolargeometrie impliziert der Bildpunkt x_1 im Bild 2 eine Epipolarlinie, auf der der Bildpunkt x_2 liegt. Der umgekehrte Fall gilt analog. Skizze: Wikipedia – Die freie Enzyklopädie.

damentalmatrix $\mathbf{F}$. Sie ist nicht auf die Verwendung nur einer Kamera beschränkt. Es gilt

$$\mathbf{F} \doteq (\mathbf{K}'^{-1})^T \boldsymbol{R}' \mathbf{S}_T \boldsymbol{R}''^{-1} \mathbf{K}''^{-1},$$

wobei die Matrizen K' und K'' die Kalibrierung, $\boldsymbol{R}'$ und $\boldsymbol{R}''$ die Rotation der Kameras sowie die schiefsymmetrische Matrix $\mathbf{S}_T$ die Translation zwischen den Projektionszentren beschreiben. Damit kann die Koplanarität (mit homogenen Vektoren $\mathbf{x}_1, \mathbf{x}_2$) als Bedingung für die Homologie der Bildpunkte einfach dargestellt werden:

$$\mathbf{x}_1^T \mathbf{F} \mathbf{x}_2 = 0.$$

Die Epipolargeometrie eignet sich zur Unterstützung der Bildzuordnung. Wenn ein markanter Punkt im ersten Bild lokalisiert ist, wird bei bekannter Epipolargeometrie der Suchbereich im zweiten Bild auf die so genannte Epipolarlinie eingeschränkt. Diese entsteht, indem der Bildpunkt im einen Bild mit den zwei Projektionszentren eine Ebene bildet, mit der die Bildebene des anderen Bildes geschnitten wird.

Essentielle Matrix

Eine Spezialisierung der Fundamentalmatrix ist die essentielle Matrix $\mathbf{E}$ $(\boldsymbol{R}'\mathbf{S}_T\boldsymbol{R}''^{-1})$. Diese ergibt sich, wenn mittels der inneren Orientierung normierte Bildkoordinaten verwendet werden, bei denen der Ursprung eines kartesischen Koordinatensystems im Hauptpunkt des Bildes liegt und folgende Bedingungen (neun kubische Gleichungen) erfüllt sind:

$$\mathbf{E}\mathbf{E}^T\mathbf{E} - \frac{1}{2} Spur(\mathbf{E}\mathbf{E}^T)\mathbf{E} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Der zur eindeutigen Lösung der Gleichungen notwendige Minimalsatz umfasst lediglich fünf Korrespondenzen (Nistér, 2003). Aus Sicht der geometrischen Räume beschreibt die Fundamentalmatrix einen projektiven Raum und die essentielle Matrix einen ähnlichen Raum.

2.2 3D Ähnlichkeitstransformation

Dreidimensionale, euklidische Vektorräume über einem beliebigen Zahlenkörper sind isomorph, d.h. sie können grundsätzlich umkehrbar eindeutig aufeinander abgebildet werden. Eine mögliche Abbildung ist die bijektive 3D Ähnlichkeitstransformation. Sie besteht aus Translation, Maßstab

und Rotation und kann durch drei Punktkorrespondenzen eindeutig bestimmt werden (siehe Abschnitt 5.1). Für zwei euklidische Vektorräume V_1, V_2 über $\mathbb{R}^3$ existiert somit ein Skalar $m \in \mathbb{R}$, eine Rotationsmatrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ und Skalare $t_X, t_Y, t_Z \in \mathbb{R}$, so dass für korrespondierende Vektoren $(X,Y,Z) \in V_1, (x,y,z) \in V_2$ gilt:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = m\boldsymbol{R} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_X \\ t_Y \\ t_Z \end{pmatrix}.$$

2.3 Robuste Parameterschätzung und Bündelausgleichung

In der Praxis weist jedes Messverfahren Fehler auf. Abhängig von Aufgabe und Methodik kann die Größe des Fehlers vernachlässigbar oder sehr gravierend sein. Es wurde eine Vielzahl von Verfahren entwickelt, um auch bei groben Ausreißern verlässlich zu genauen Resultaten gelangen zu können.

Für diese Arbeit wurden zum Zwecke der 3D Rekonstruktion zwei gängige Methoden kombiniert, nämlich RANSAC und robuste Kleinste-Quadrate-Schätzung. RANSAC ist eine Methode zur Schätzung eines Modells innerhalb einer Reihe von Messwerten auch mit groben Ausreißern (Fischler & Bolles, 1981). Dabei wird ein minimaler Messwertesatz zur Bestimmung einer Instanz des Modells zufällig ausgewählt und die Instanz des Modells mit den übrigen Messwerten verifiziert. Die Messwerte, die mit der Instanz des Modells korrespondieren werden hierbei als Inlier bezeichnet. Angewandt auf die relative Orientierung eines Bildpaares im kalibrierten Fall bedeutet dies, dass Instanzen des Modells auf der Grundlage von fünf Punktkorrespondenzen, welche den minimalen Datensatz zur eindeutigen Bestimmung der essentiellen Matrix darstellen, erzeugt werden.

Die Methode der kleinsten Quadrate ist das mathematische Standardverfahren zur Ausgleichungsrechnung und geht auf Carl Friedrich Gauß zurück. Sie besteht darin, die Parameter einer gegebenen Funktion so zu bestimmen, dass die Summe der Quadrate der Abweichungen der Funktion von den beobachteten Punkten minimiert wird. Voraussetzung ist, dass die Fehler der Eingangsdaten bezüglich der Funktion normalverteilt sind und keinerlei Systematik aufweisen.

Somit können mittels RANSAC geeignete Hypothesen gefunden und mittels Kleinster-Quadrate verbessert werden. Unter alleiniger Verwendung dieser Methoden können Ausreißer, die deutlich über der statistisch zu erwartenden Abweichung liegen, dazu führen, dass auch grob falsche Modelle als gute Lösungen erkannt werden. Des Weiteren ist es nicht sinnvoll, zur Bewertung eines

gefundenen Modells allein die Zahl der Inlier zu verwenden. Aus diesem Grund konzipierte Torr (1998) das Geometric Robust Information Criterion (GRIC):

$$GRIC = \sum p(e_i^2) + \lambda_1 dn + \lambda_2 k, \text{ mit } 1 \leq \lambda_1 \leq 2 \text{ und } \lambda_2 > 2.$$

Auf Grundlage der Anpassungsfehler aus der Kleinste-Quadrate-Zuordnung ($p(e_i^2)$), der Dimension der Bedingung (d), der Zahl der Inlier (n) und dem Freiheitsgrad des Modells (k) können mit diesem Kriterium geometrische Modelle auch bezüglich der erzielten Genauigkeit bewertet werden, so dass das Aussortieren von falschen Lösungen und Erkennen der korrekten Lösung verbessert wird.

Bei der Bündelausgleichung wird die Methode der Kleinsten Quadrate auf die Optimierung der Sehstrahlenbündel einer Szene angewandt, welche von mehreren Kameras beobachtet wird. Dabei werden gleichzeitig die Positionen der Objektpunkte und Orientierungen der Kameras sowie evtl. deren Kalibrierparameter angepasst, so dass die gewichtete Summe der verbleibenden Fehler, quadratisch für alle Beobachtungen minimal wird.

2.4 Gaußfunktion zur Glättung von Bildfunktionen

Zur Glättung des Bildinhaltes werden oft Gauß-Filter verwendet. Damit kann das Bildrauschen vermindert werden. Kleinere Strukturen gehen verloren, gröbere Strukturen bleiben dagegen erhalten. Die 2D Impulsantwort der Gaußfunktion lautet:

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

Die Glättung einer Bildfunktion $I(x,y)$ erfolgt durch Faltung:

$$(G(\cdot,\cdot;\sigma) * I)(x,y).$$

Dabei entspricht σ der Standardabweichung der Gaußfunktion. Spektral entspricht die Glättung einem Tiefpassfilter. Durch Gaußfilterung (siehe Abb. 2.4) kann die Bildfunktion in Abhängigkeit von σ so verändert werden, dass das geglättete Bild näherungsweise dem entspricht, das man bei einem größerem Abstand der Kamera vom Objekt erhalten würde. Damit ist auch die entsprechende Skalierung eines Bildes möglich. Mathematisch ist auch die Verkürzung des Abstands zur Szene definiert, was einer Hochskalierung entspräche, allerdings ist dies für ein Bild nicht sinnvoll, da die erforderliche Information zur Darstellung der Szene nicht vorhanden ist. Die Faltung

mit der Gaußfunktion liefert in Abhängigkeit von σ eine Maßstabsraumebene M der Bildfunktion I:

$$M(x,y;\sigma) = (G(\cdot,\cdot;\sigma) * I)(x,y).$$

Die Maßstabsraumebenen haben bei der Belegung von σ mit Zweierpotenzen einen identischen Informationsgehalt wie Stufen einer Bildpyramide (Koenderink, 1984). Dabei entspricht die Gaußfunktion mit $\sigma = 0$ per Definition der Identität. Die Stufen λ einer Bildpyramide und die Standardabweichung σ stehen im Zusammenhang:

$$\lambda = ld(2\sigma^2).$$

Durch diese Beziehung kann von der Standardabweichung σ auf die entsprechende Skalierung eines Bildes geschlossen werden.

Abbildung 2.4: Die Bilder veranschaulichen die Wirkungsweise der Gaußfunktion zur Glättung der Bildfunktion. Von links oben nach rechts unten sind das Originalbild und die geglätteten Bilder mit $\sigma = 1,2,4,8,16$ dargestellt. Bilder: Wikipedia – Die freie Enzyklopädie.

Kapitel 3

Stand der Forschung

3.1 Extraktion markanter Punkte

Die Bestimmung von Korrespondenz, oft in Form von homologen Punkten, ist eine grundlegende Voraussetzung zur Orientierung von Bildern. Für eine automatisierte Bestimmung sind zahlreiche Operatoren zur automatisierten Extraktion von markanten Punkten aus digitalen Bildern entwickelt worden. Die extrahierten Punkte sollen charakteristisch sein, damit sie zuverlässig eine korrekte Zuordnung korrespondierender Punkte in verschiedenen Bildern ermöglichen. Zu den häufig verwendeten Punkt Operatoren zählen der Förstner-Operator (Förstner & Gülch, 1987), der Harris-Operator (Harris & Stephens, 1988) und der Scale-Invariant-Feature-Transform (SIFT) Operator (Lowe, 2004).

Der Förstner-Operator (Förstner & Gülch, 1987) verwendet die Autokorrelation von Bildausschnitten als Grundlage zur Bestimmung markanter Punkte. Es wird ausgenutzt, dass die Inverse der Autokorrelationsmatrix (Strukturtensor) der Kovarianzmatrix entspricht. Letztere gibt Aufschluss über die Lokalisierbarkeit eines markanten Punktes im Sinne einer Aufenthaltswahrscheinlichkeit. Die Eigenwerte der Kovarianzmatrix beschreiben die Genauigkeiten, mit denen ein markanter Punkt in den Hauptrichtungen bestimmt werden kann. Kleine, in etwa gleich große Eigenwerte lassen auf einen in alle Richtungen gut lokalisierbaren markanten Punkt schließen. Der Förstner-Operator ist auch als Kantendetektor einsetzbar: Kanten führen zu deutlich unterschiedlich großen Eigenwerten (Förstner, 1994). Auch der Harris-Operator (Harris & Stephens, 1988), der auch Plessy-Punkt-Detektor genannt wird, verwendet die Autokorrelation von Bildausschnitten als Grundlage. Für die Detektion eines markanten Punktes wird jedoch lediglich der Rang der Autokorrelationsmatrix betrachtet.

Für den Förstner- und den Harris-Operator ist es möglich, die Orientierung der Fehlerellipse für die Normierung der Rotation des Punktes zu verwenden. Diese liefert aufgrund ihrer Symmetrie zwei, um π gedrehte Lösungen. Eine eindeutige Bestimmung ist unter Hinzunahme weiterer Information, wie beispielsweise der Polarisation (hell-dunkel) möglich. Der SIFT-Operator benutzt ein Histogramm über die Pixelunterschiede der jeweiligen Bildregion, um die Orientierung eines Punktes in der Bildebene über den gesamten Wertebereich des Vollkreises eindeutig zu bestimmen. Wenn markante Punkte hinsichtlich ihrer Orientierung in der Bildebene normalisiert werden, ist es möglich, festzustellen, wie zwei markante Punkte gegeneinander in der Bildebene gedreht sind. Sind sowohl die Zuordnung als auch die bestimmten Orientierungen in der Bildebene korrekt, dann kann aus dem Unterschied der Orientierungen auf den Winkel geschlossen werden, um den die Bilder um die Hauptachse rotiert sind.

Die Ergebnisse von Förstner- und Harris-Operator sind nur dann gut zuzuordnen, wenn eine Szene in einem möglichst gleich bleibenden Maßstab abgebildet wird. Dies schränkt die Nutzung und damit die möglichen Aufnahmekonfigurationen stark ein. Mit (Lindeberg, 1994), wurde ein Konzept präsentiert, mit dem maßstabsinvariante Bildzuordnung ermöglicht wird. Es verwendet die Laplace-Gaußfunktion L zur Normierung des Maßstabs von markanten Punkten.

Die Differenz $D(x,y,\sigma)$ zwischen zwei um einen konstanten Faktor k verschiedenen Gaußebenen $G(x,y;\sigma)$ und $G(x,y;k\sigma)$ mit $k \in \mathbb{R}$ ergibt sich zu:

$$D(x,y;\sigma) = (G(x,y;k\sigma) - G(x,y;\sigma)) * I(x,y) = L(x,y;k\sigma) - L(x,y;\sigma).$$

Die bezüglich des Maßstabs normierte Laplace-Gaußfunktion $\sigma^2\nabla^2 G$ kann über die Differenzen der Gaußebenen näherungsweise bestimmt werden. Für kleine k gilt für die partielle Ableitung nach σ folgende Beziehung zum Differenzenquotienten:

$$\sigma\nabla^2 G = \frac{\partial G}{\partial \sigma} \approx \frac{G(x,y,k\cdot\sigma) - G(x,y,\sigma)}{(k\cdot\sigma - \sigma)} = \frac{G(x,y,k\cdot\sigma) - G(x,y,\sigma)}{(k-1)\cdot\sigma}.$$

Daraus folgt:

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

Um Punkte und einen Maßstab zu bestimmen, wird nach Maßstabsraumextremwerten gesucht. Diese Methode ist für Blob-Strukturen geeignet. Blobs sind kleine helle oder dunkle Flecken im Bild. Dafür werden für jede Oktave eine bestimmte Anzahl an Gauß-Ebenen und die daraus resultierenden Differenz-Ebenen bestimmt. Eine Oktave entspricht jeweils der halben Bildgröße ihres

Vorgängers. Typischerweise werden drei Differenz-Ebenen pro Oktave verwendet, was einen guten Kompromiss zwischen Rechenaufwand und der Anzahl an gut zuzuordnenden Punkten ergibt. Es ist aber grundsätzlich möglich, deutlich mehr Ebenen zu verwenden, und abhängig von der Anwendung auch sinnvoll.

Zum Test, ob an der Stelle (x, y, σ) ein Extremwert vorliegt, wird der Wert mit allen acht umliegenden Pixeln auf der Bildebene σ und mit den jeweils neun der höher- und tieferliegenden Bildebene verglichen (siehe Abb. 3.1). Damit werden drei verschiedene Differenz-Ebenen verwendet. Ein Extremwert liegt vor, wenn der Intensitätswert des Kandidaten jeweils höher (Maximum) bzw. niedriger (Minimum) ist, als alle 26 verglichenen umliegenden Werte.

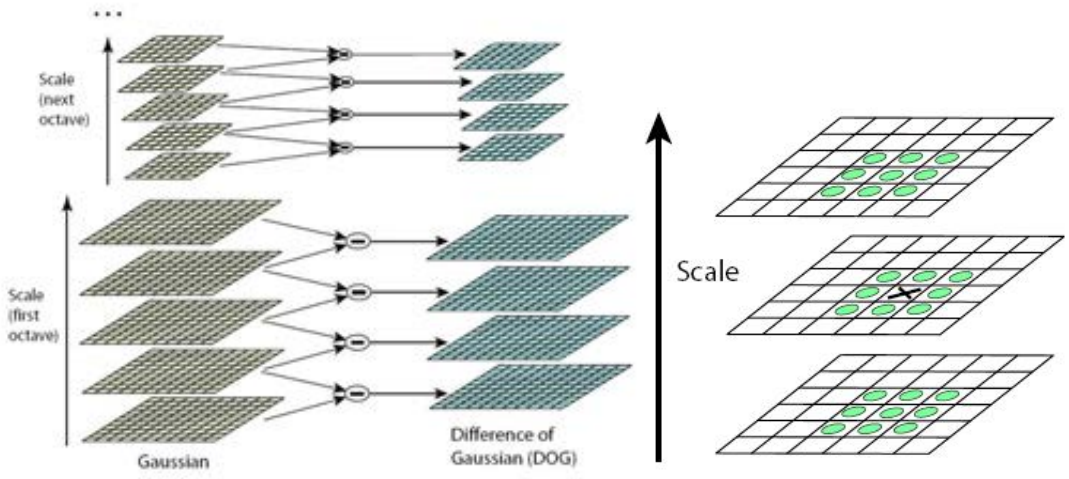


Abbildung 3.1: Links: Berechnung der Differenzen von Gauß-Ebenen. Rechts: Bestimmung eines Maßstabsraumextremwerts durch Vergleich mit 26 umliegenden Pixeln. Skizze: (Lowe, 2004).

Der SIFT-Operator ist ein deskriptorbasierter Blob-Operator von Lowe (2004). Er ist invariant gegen Translation, Rotation um die Hauptachse und Maßstabsänderungen. Er verwendet die oben vorgestellte Methodik zur Bestimmung von Maßstabsraumextremwerten.

Der 128 dimensionale SIFT Deskriptor wird aus einem 16 $\times$ 16 Pixel großen Bildausschnitt bestimmt, nachdem ein markanter Punkt (x, y, o, σ) gefunden wurde, wobei (x, y) für die Lokalisierung, o für die Rotation in der Bildebene (Bestimmung siehe oben) und σ für die Gaußebene, auf der der Maßstabsraumextremwert gefunden wurde, stehen. Auf der Gaußebene σ wird ein Ausschnitt, welcher hinsichtlich der bestimmten Rotation o normiert wird, in 16 Teilausschnitte der Größe 4 $\times$ 4 Pixel unterteilt (siehe Abb. 3.2, links). Die Gradienten jedes 4 $\times$ 4 großen

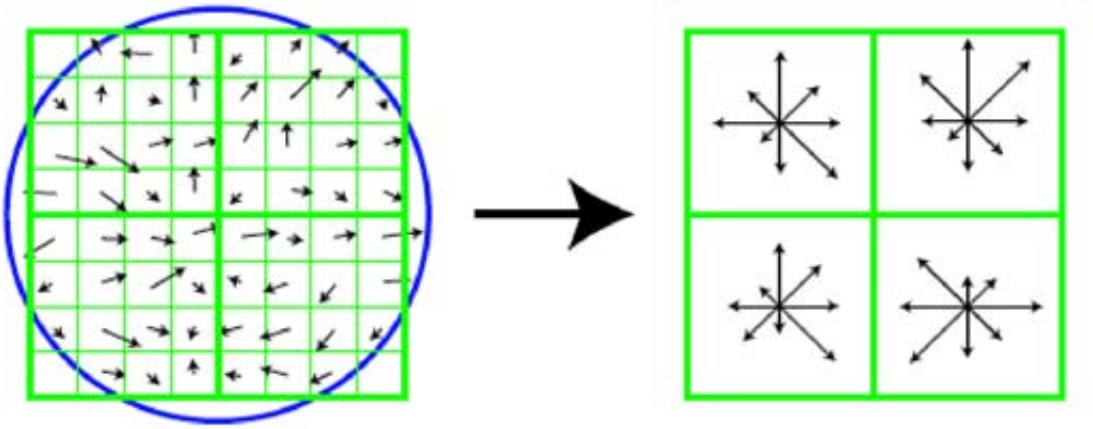


Abbildung 3.2: Definition des SIFT-Deskriptors. Skizze: (Lowe, 2004).

Teilausschnitts werden auf acht Hauptrichtungen projiziert. Dabei wird der Betrag eines Gradienten abhängig von seiner Richtung und unter Berücksichtigung einer Gewichtung mittels der Gaußfunktion, auf die Hauptrichtungen addiert (siehe Abb. 3.2, rechts). Die acht Hauptrichtungen werden durch einen acht-dimensionalen Vektor repräsentiert, was bei 16 Teilausschnitten zu einem $16 \times 8 = 128$ dimensionalen Gesamtvektor, dem Deskriptor, führt. Durch eine Gewichtung werden Gradienten im Zentrum des Gesamtausschnitts stärker berücksichtigt, als am Rande. Dadurch verändert sich ein Deskriptor bei kleiner Verschiebung des Gesamtausschnitts nur wenig und eine ungenaue Lokalisierung eines SIFT-Features kann im hohen Maße kompensiert werden.

Für den SIFT-Operator wurde von Wu (2007) eine leistungsstarke, NVIDIA GPU basierte Umsetzung entwickelt und frei zur Verfügung gestellt. Diese Implementierung ermöglicht eine echtzeitnahe Anwendung.

Der SFOP-Operator von Förstner *et al.* (2009) wurde mit der Zielsetzung entwickelt, auch für Bilder, die z.B. weitgehend eintönige Fassaden zeigen, eine große Anzahl an maßstabsnormierten markanten Punkten extrahieren und zuordnen zu können. Dies ist mit dem SIFT-Operator unter diesen Umständen grundsätzlich problematisch. Für den SFOP-Operator wurde das Konzept von Bigün (1990) zur Bestimmung von Blob-, Knoten- und Kreisstrukturen verwendet.

3.2 Bestimmung von Punktkorrespondenzen

Zur Bestimmung von korrespondierenden Punkten wurde in Photogrammetrie und Computer Vision eine Vielzahl von Methoden entwickelt. Dabei kann grundsätzlich zwischen Methoden unterschieden werden, welche Bildausschnitte (Kreuzkorrelation und affine Kleinste-Quadrate- Zuordnung) und welche charakterisierende Deskriptoren (SIFT) verwenden.

Eine etablierte Bildausschnitt-basierte Zuordnungsmethode beruht auf normierter Kreuzkorrelation (Normalized Cross Correlation – NCC). Dabei werden um extrahierte Punkte Bildausschnitte von einer Größe von typischerweise 9×9 bis 15×15 Pixel ausgewählt. Für zwei zu vergleichende Bildausschnitte P_1 und P_2 werden jeweils die Varianzen $Var(P_1)$ und $Var(P_2)$ und die Kovarianz $CoVar(P_1, P_2)$ der Intensitätswerte bestimmt. Der normierte Kreuzkorrelationskoeffizient ρ_{P_1,P_2} ergibt sich dann als:

$$\rho_{P_1,P_2} = \frac{CoVar(P_1,P_2)}{\sqrt{Var(P_1) \cdot Var(P_2)}}. \tag{3.1}$$

mit: $-1 \leq \rho_{P_1,P_2} \leq 1$.

Normierte Kreuzkorrelation ist aufgrund der Einbeziehung von Mittelwerten und Varianzen invariant bezüglich gleichmäßiger Veränderungen von Helligkeit und Kontrast.

Zur Bildzuordnung für SIFT-Merkmale wurde von Lowe (2004) ein Verfahren entwickelt, welches das aufwändige Austesten von $n \times m$ Punkt-Zuordnungsmöglichkeiten in zwei Bildern unnötig macht. Dem Zuordnungsverfahren liegt ein sortierter 128-dimensionaler k-d Baum zu Grunde, welcher aus ca. 40.000 SIFT-Punkten, extrahiert aus Trainingsbildern, abgeleitet wurde. In diesen sortierten Referenz k-d Baum werden SIFT-Deskriptoren einsortiert. Sind alle Deskriptoren eines Bildpaares einsortiert, werden für die Zuordnung Algorithmen zur Suche des nächsten Nachbars im k-d Baum verwendet. Die Stärke dieser Vorgehensweise besteht in der geringen Laufzeit. Denn obgleich hinsichtlich der Zuverlässigkeit Abstriche gemacht werden müssen, wird der Aufwand verglichen mit dem vollständigen Prüfen aller $n \times m$ Zuordnungsmöglichkeiten deutlich reduziert: Die Anzahl an SIFT-Punkten in einem Bild beläuft sich auf typischerweise einige hundert bis einige tausend.

Bestimmung der Rotation eines Bildes um die Hauptachse

Basierend auf der Bestimmung der Orientierung der markanten Punkte in der Bildebene (siehe

Abschnitt 3.1) kann für zwei perspektivisch gegenseitig nicht extrem verzerrte Bilder durch statistische Auswertungen die Rotation um die Hauptachse zuverlässig bestimmt werden (Mayer, 2007a,b). Hierbei wird berücksichtigt, dass die für einzelne Punkte bestimmte Orientierung je

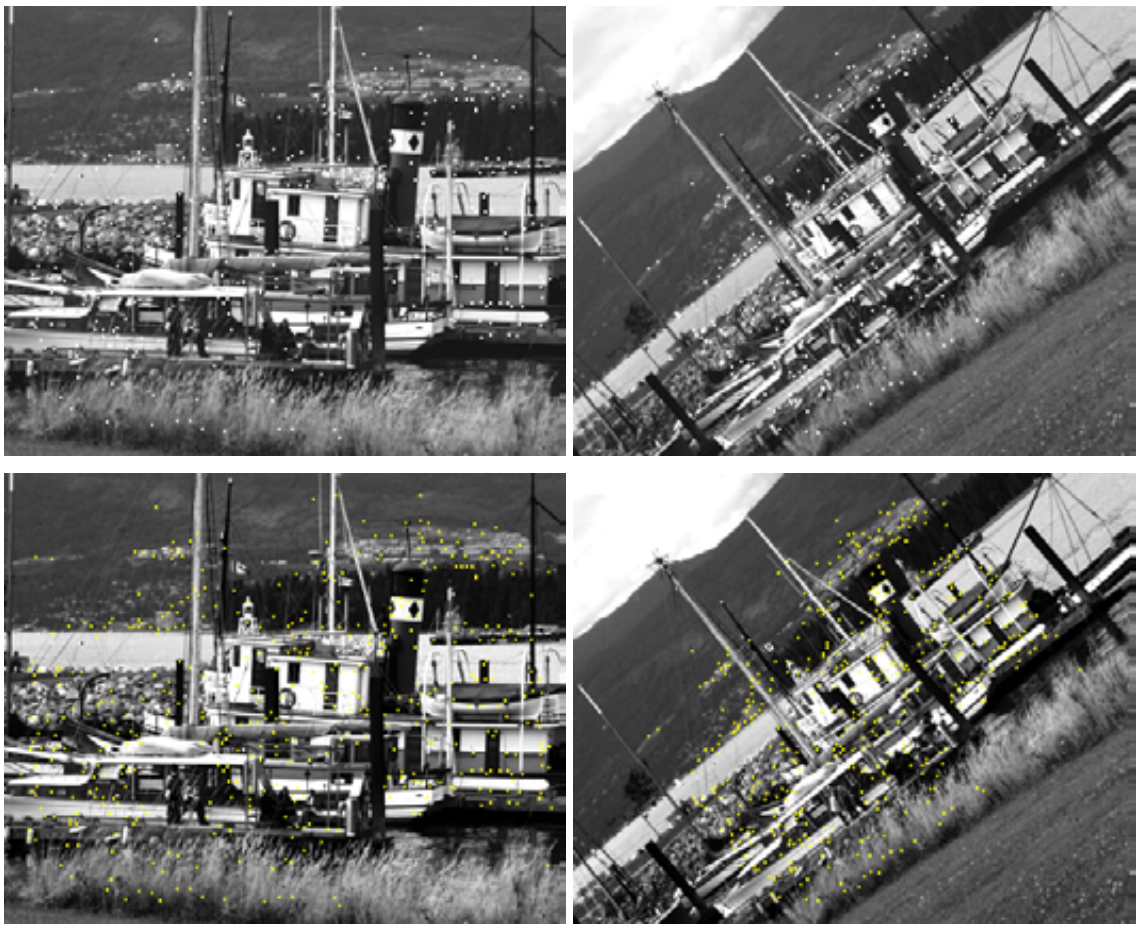


Abbildung 3.3: Oben: Ergebnis einer Zuordnung auf der Grundlage normierter Kreuzkorrelation (NCC) ohne Einbeziehung der Rotation um die Hauptachse: 327 korrekte Zuordnungen (weiß) von 412 insgesamt (79% korrekt). Unten: Ergebnis mit Einbeziehung der Rotation um die Hauptachse (40,5°): 437 korrekte Zuordnungen (gelb) von 511 insgesamt (86% korrekt).

nach Beschaffenheit der zuzuordnenden Bilder einen Fehleranteil aufweist. Um trotzdem für alle zuzuordnenden Punkte die Rotation um die Hauptachse korrekt berücksichtigen zu können, wird zunächst eine Bildzuordnung auf Grundlage der bezüglich der Rotation normierten markanten Punkte durchgeführt. Für alle zuzuordnenden Punkte kann somit ein Differenzwinkel bezüglich

der Orientierung in der Bildebene und je nach Zuordnungsverfahren ein Ähnlichkeitsmaß bzw. Matchingscore bestimmt werden. Dementsprechend kann jeder Differenzwinkel einem Intervall zugeordnet werden. Eine sinnvolle Intervallgröße beträgt 5 bis 10°. Es kann dann für jedes Intervall gezählt werden, wie viele Zuordnungen auftreten. Dies entspricht einem Histogramm. Die Rotation um die Hauptachse um den Winkel α wird durch das Intervall mit den meisten Zuordnungen bestimmt und für einen zweiten Zuordnungsvorgang verwendet. Bei diesem werden alle markanten Punkte im ersten Bild nicht und im zweiten Bild alle um den Winkel α rotiert. Dies führt, wie in Abb. 3.3 dargestellt, zu verbesserten Resultaten. Für diese Demonstration wurden zwei Bilder des quasi Benchmarktestdatensatzes der Robotik Gruppe der Universität Oxford (Visual Geometry Group, 2003) verwendet. Die zugehörige Homographie ermöglicht die Verifikation der auf der Grundlage von NCC bestimmten Punktzuordnungen.

Affine Kleinste-Quadrate-Zuordnung

Je größer die Blickpunktänderung zwischen zwei zuzuordnenden Bildern ist, desto größer kann die perspektive Verzerrung der abgebildeten Objekte werden. In der Szene übereinstimmende Bildregionen sehen damit im einen Bild deutlich verschieden aus als im anderen. Perspektive Verzerrung erschwert die Bildzuordnung daher erheblich. Eine korrekte Zuordnung ist je nach Methode durch die Größe der Blickpunkt- und Maßstabsänderung begrenzt. Eine ganze Reihe Zuordnungsverfahren verwenden die affine Transformation, um dieses Problem anzugehen (Mikolajczyk *et al.*, 2005). Obgleich für die exakte Beschreibung der perspektiven Verzerrung einer ebenen Fläche eine projektive Transformation erforderlich ist, erweist sich die affine Transformation in vielen Fällen als hinreichend genau, sofern die Ausdehnung der zuzuordnenden Regionen klein gewählt wird. Auf diese Weise kann gleichzeitig auch eine fehlende Ebenheit der Szene z.T. kompensiert werden.

Die affine Transformation $\boldsymbol{A} :: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ besteht aus einer Translation (t) und Rotation in der Ebene ($\boldsymbol{R}$), Maßstab (M), sowie Scherung (S) und Dehnung (D). Sie wird beschrieben durch:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \boldsymbol{R} \begin{pmatrix} 1 & S \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} M & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix},$$

$$\text{mit } \boldsymbol{R} = \begin{pmatrix} \cos\Theta & \sin\Theta \\ -\sin\Theta & \cos\Theta \end{pmatrix}.$$

Unter Verwendung homogener Koordinaten folgt:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & t_x \\ A_{21} & A_{22} & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \mathbf{A} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}.$$

Durch geeignete Wahl der Transformationsparameter kann die Auswirkung einer Drehung einer Ebene im Raum approximiert werden (siehe Abb. 3.4). Das transformierte Bild gleicht stark der Perspektive aus einem anderen Blickwinkel. Auf diese Weise können die durch perspektive Verzerrung verursachten Unterschiede in größeren Teilen korrigiert werden.

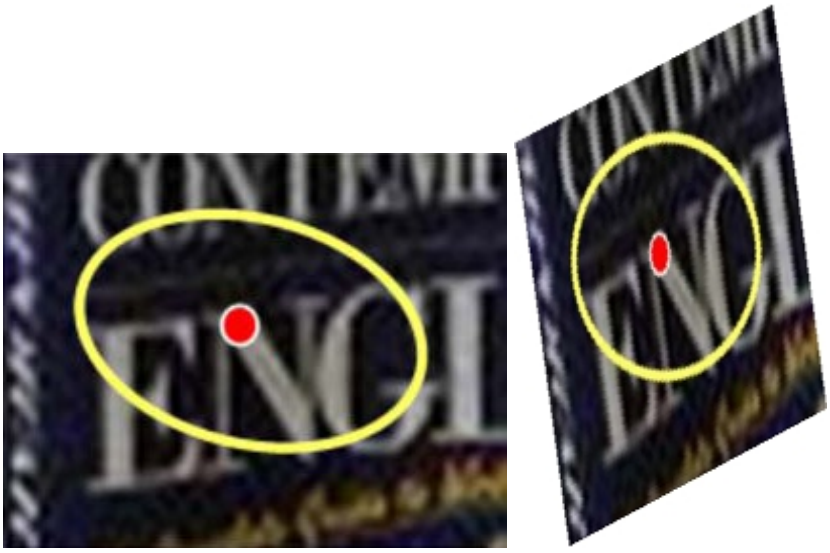


Abbildung 3.4: Auswirkung von affiner Transformation auf Bildausschnitte. Der Originalbildausschnitt (links) wurde in x-Richtung um 50% gestaucht (Dehnung) und um 15° gegen den Uhrzeigersinn um die Hauptachse gedreht. Anhand der dann kreisförmigen gelben Markierung im transformierten Bildausschnitt (rechts) wird die Veränderung der Bildregion deutlich. Bild: (Mikolajczyk *et al.*, 2005)

In Verbindung mit den im Abschnitt 2.3 erläuterten Methoden zur Ausgleichung ist es möglich, die Transformationsparameter für zuzuordnende Bildregionen zu schätzen (Grün, 1985). Hierbei ist zu beachten, dass das bestimmte Optimum lokal ist, d.h. von der Wahl von Näherungswerten abhängig ist.

3.3 Bildzuordnung für große Blickwinkeländerungen

In diesem Abschnitt werden Zuordnungsverfahren für Punkte betrachtet, die dafür konzipiert wurden, möglichst robust gegen große Blickwinkeländerungen zu sein.

Im Folgenden werden einige wichtige Aspekte von bisherigen Arbeiten auf Grundlage von (Mikolajczyk *et al.*, 2005) erläutert. Die von Mikolajczyk *et al.* (2005) präsentierten Ergebnisse werden von sehr vielen Autoren zum Vergleich herangezogen. Ein dort beschriebener Ansatz ist der Maximally Stable Extremal Regions (MSER) Detektor (Matas *et al.*, 2002). Er verwendet Bildregionen, die durch lokale Binarisierung bestimmt werden. Für die so erhaltenen Bildregionen erfolgt eine Normalisierung aller sechs Parameter der affinen Transformation. Die Autoren erwarteten, dass diese Regionen stabil gegen perspektive Veränderung, Maßstabsunterschiede und gleichmäßige Änderung der Lichtverhältnisse sind. Die Ergebnisse der quasi Benchmark-Testdatensätze (Visual Geometry Group, 2003) belegen jedoch, dass dieser Operator nur wenig robust gegen die genannten Einflüsse ist.

Der Harris-Affine-Operator verwendet Harris-Punkte (siehe Abschnitt 3.1) in Verbindung mit der Methode von Lindeberg (1994, 1998) zur Bestimmung von Maßstabsraumextremwerten. Die Form eines Bildausschnitts wird aus den Eigenwerten der Matrix der zweiten Momente abgeleitet und iterativ zur Kreisform normalisiert. Der Hesse-Affine-Operator funktioniert nach demselben Prinzip, aber als Blob Operator. Die Ergebnisse bzgl. Robustheit gegen Maßstabsunterschiede und Blickwinkeländerungen sind nicht zufriedenstellend. Morel & Yu (2009) diskutieren die Ursachen und kommen zu dem Schluss, dass bei der Normalisierung Bildregionen je nach Beschaffenheit auch hochskaliert werden, was nicht sinnvoll ist (siehe Abschnitt 2.4). Hinsichtlich Blickwinkeländerungen ist eine zuverlässige Zuordnung nur in geringfügig höherem Maße möglich, als dies bei der SIFT-Zuordnung der Fall ist.

Morel & Yu (2009) veröffentlichten mit Affine-SIFT (ASIFT) einen Ansatz, welcher dafür konzipiert wurde, möglichst robust gegen Blickwinkeländerungen und Maßstabsunterschiede zu sein. ASIFT verwendet die affine Transformation zur Korrektur der perspektiven Verzerrung (siehe Abb. 3.4). Dabei werden Neigung und Schwenkung simuliert und nicht wie die übrigen Parameter normiert. Die Bestimmung des Maßstabsfaktors und die Normierung von Translation und Rotation um die Hauptachse erfolgt durch den SIFT-Operator. Die Bilder werden bezüglich der Neigungs- und Schwenkungsparameter über einen festgelegten Suchbereich in Stufen affin transformiert, wobei anisotrope Gauß-Glättung Aliasing Effekte gering hält. Der Vergleich erfolgt über SIFT-Zuordnung. Dann wird geprüft, ob für eine affine Transformation eine hinreichend große Anzahl an wahrscheinlich korrekten Zuordnungen gefunden wurde. Auf diese Weise werden die signifikant besten affinen Transformationen bestimmt.

Unter Verwendung der signifikanten affinen Transformationen und unter Einbeziehung der geometrischen Plausibilität wird der Zuordnungsvorgang auf einer höheren Bildauflösung wie-

derholt. Mit dieser Vorgehensweise wurden für die quasi Benchmark-Testdatensätze (Visual Geometry Group, 2003) sehr gute Ergebnisse erzielt. Zu beachten ist, dass die affine Transformation lediglich eine globale Dehnung zulässt und deshalb nur für eine lokale Angleichung der perspektiven Verzerrung geeignet ist. Deshalb ist im Allgemeinen zu erwarten, dass für ein Bildpaar eine ganze Reihe an Transformationen erforderlich ist, auch wenn die Szene nur aus einer Ebene besteht. Der Rechenaufwand nimmt mit der Anzahl der ermittelten affinen Transformationen zu, da im zweiten Zuordnungsschritt alle signifikanten Transformationen angewandt werden. Damit hängt der Rechenaufwand dieser Methode von der Beschaffenheit einer Szene ab.

3.4 Orientierung großer Bildverbände

Pollefeys *et al.* (2004) präsentierten bereits im Jahr 2004 einen umfassenden Ansatz zur 3D Modellierung urbaner Szenen, welcher auf Bildern handelsüblicher Kameras beruht. Es wurde gezeigt, dass es mittels Punkt-Operatoren und Kreuzkorrelation als Zuordnungsverfahren möglich ist, Bilder zu orientieren. Dafür wurde die Epipolargeometrie durch die Fundamentalmatrix (siehe Abschnitt 2.1) bestimmt. Des Weiteren wurde ein Ansatz zur dichten Tiefenschätzung vorgestellt. Mit (Pollefeys *et al.*, 2008) wurde ein echtzeitfähiger Ansatz zur georeferenzierten 3D Rekonstruktion bebauter Gebiete präsentiert. Dieser verwendet Videos in Verbindung mit GPS und INS Information.

Photosynth ist ein Projekt von Microsoft zur Orientierung von und 3D Rekonstruktion aus beliebig zu Stande gekommenen Bildverbänden. Es basiert auf einer Web-basierten Client- / Server Architektur, die es jedermann ermöglicht, Bilder einer Szene einzusenden und der 3D Rekonstruktion der Szene hinzuzufügen. Die Bilder werden ohne weitere Zusatzinformation serverseitig ausgewertet und integriert. Neben den 3D Punkten werden über den Client jeweils die orientierten Bilder dargestellt, die am besten zum momentanen Blickpunkt und -richtung passen (siehe Abb. 3.5). Bis zur Einstellung der 3D Visualisierung in Bing Maps (siehe Abschnitt 1.2) konnten Benutzer Photosynth dazu verwenden, dessen Datenbasis zu erweitern.

Agarwal *et al.* (2009) präsentierten mit „Building Rome in a Day“ einen Ansatz zur Orientierung sehr großer Bildverbände, welcher mehrere bekannte leistungsstarke Verfahren kombiniert. Als Grundlage wurde das Open Source Verfahren Bundler (Snavely, 2010) verwendet, welches 3D Rekonstruktion auf der Grundlage unsortierter Bildsequenzen mittels der von Schaffalitzky & Zisserman (2002) vorgestellten Methode ermöglicht. Dabei werden der SIFT-Operator für Punktextraktion und Zuordnung (siehe Abschnitte 3.1 und 3.2) und eine auf dem Levenberg-Marquardt

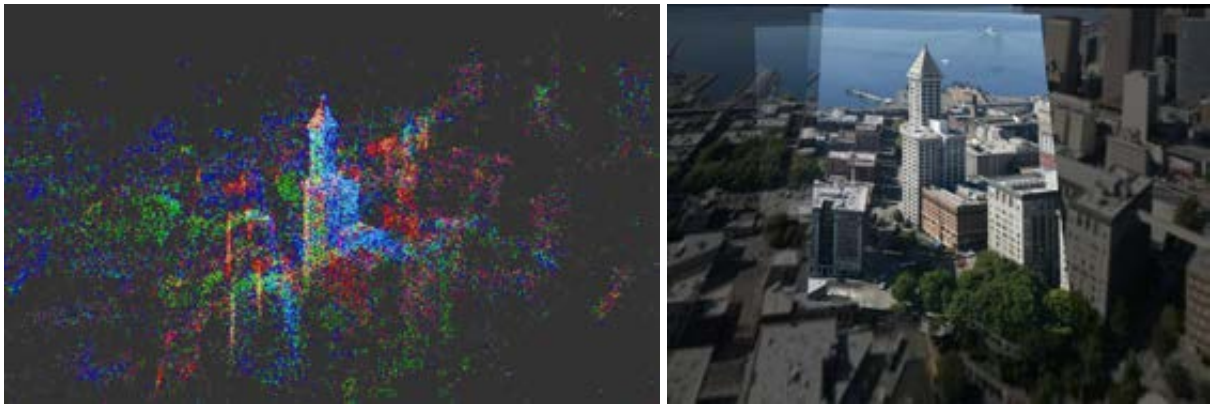


Abbildung 3.5: Der Smith Tower (Seattle, USA) als 3D Rekonstruktion in Photosynth (Microsoft, 2011). Links: 3D Punktwolke, rechts: Blick auf die Szene aus der Perspektive eines orientierten Bildes.

Verfahren basierende Bündelausgleichung (Lourakis & Argyros, 2009) angewendet. Diese Kombination ermöglicht die Orientierung sehr großer, zufällig zu Stande gekommener Bildverbände aus dem Internet (siehe Abb. 3.6). Die Verwendung von SIFT-Zuordnung zur Bildzuordnung führt zu wenig Toleranz bzgl. größerer Blickwinkeländerung (siehe Abschnitt 3.2). Daher setzt dieser Ansatz eine hohe Aufnahmedichte voraus, was vor allem an stark besuchten Tourismuszielen gegeben ist. Unter Nutzung von Hochleistungsrechensystemen ist die Orientierung tausender Bilder in wenigen Tagen möglich.

Abbildung 3.6: Ergebnis des Projekts „Building Rome in a Day“ für das Kolosseum, Rom (Agarwal *et al.*, 2009). Es wurden 2.106 Bilder verwendet und die 3D Rekonstruktion beinhaltet 819.242 3D-Punkte. Die Berechnungen erfolgten mittels eines Hochleistungsrechensystems.

„Building Rome on a Cloudless Day “ (Frahm *et al.*, 2010) wurde dafür konzipiert, große Teilmengen sehr großer unsortierter Bildmengen zu orientieren und eine 3D Rekonstruktion dieser Teilmengen einschließlich dichter Tiefenschätzung zur realisieren. Anders als „Building Rome in a Day“ war die Zielsetzung, Ergebnisse auf einem einzelnen PC und nicht auf einem Hochleistungsrechensystem bzw. einer Cloud (daher die Bezeichnung „Cloudless“) zu produzieren. Dafür wurde ein Großteil der Algorithmen auf der GPU implementiert. Durch den Einsatz von vier GPUs auf einem Rechner konnte eine massive Parallelisierung erreicht werden. Es wurde die in (Li *et al.*, 2008) vorgestellte Methode verwendet, um auf Grundlage von Bildern einer Internetplattform zum öffentlichen Austausch beliebiger Fotos, große, zusammenhängende Bildverbände allein auf Grundlage von Ortsnamen wie Berlin oder Rom zu generieren.

Dabei wird die Meta-Information jedes Bildes auf Angaben über Kamerakalibrierung und Positions-/ Geoinformation überprüft und ggf. in die Orientierung der Bildverbände einbezogen. Es stellte sich heraus, dass der Anteil der Bilder, die derartige Information aufwiesen, mit 40% respektive 10% höher war als erwartet. Aufgrund der Standardisierung der Digitalkameras und der Meta-Information ist zu erwarten, dass der Anteil im Internet verfügbarer Bilder, die derartige Information aufweisen, noch deutlich steigen wird. Auf die orientierten Bildverbände wird ein Verfahren zur dichten Tiefenschätzung angewendet. Die Bildzuordnung zur Orientierung der Bildverbände erfolgt mittels des SIFT-Operators. Demzufolge wird für die Bildung eines Clusters eine hohe Bilddichte vorausgesetzt, was meist nur für touristisch bedeutende Regionen gegeben ist. Durch die starke Zunahme der im Internet verfügbaren Bilder ist jedoch zu erwarten, dass die Anwendungsmöglichkeit dieses Verfahrens zunehmen wird.

Strecha *et al.* (2010) präsentierten einen Ansatz zur präzisen Orientierung sehr großer Bildverbände. Sie evaluierten dessen Leistungsfähigkeit an mehreren Beispielen (siehe Abb. 3.7). Wie in den vorangegangenen Absätzen wurden Bilder aus frei zugänglichen Internetplattformen verwendet. Der Ansatz von Strecha *et al.* (2010) basiert auf der Idee, keine großen Gesamtbildverbände zu orientieren, sondern zunächst verhältnismäßig kleine Cluster. Die Cluster werden mittels GPS-Information direkt aus den Metadaten der Bilder oder durch anderweitigen Bezug zum Weltkoordinatensystem absolut orientiert. Mittels 3D-Ähnlichkeitstransformation (siehe Abschnitt 2.2) und Ausgleichung wird eine einheitliche und präzise absolute Orientierung bestimmt. Diese Vorgehensweise hat den Vorteil, dass zur Aktualisierung nicht der komplette Bildverband neu relativ orientiert werden muss, sondern nur die betroffenen Teilbereiche. Die Autoren verweisen auch auf die Problematik, dass Bilder, die eine Szene unter deutlich verschiedenen Lichtverhältnissen

zeigen, häufig nicht relativ orientiert werden können, was für diesen Ansatz aber durch die starke Unterstützung durch GPS Information unproblematisch ist.

Abbildung 3.7: Dynamic and Scalable Large Scale Image Reconstruction – Ergebnis aus (Strecha *et al.*, 2010): 3D-Rekonstruktion von Teilen der Stadt Lausanne (Schweiz).

3.5 Grenzen bisheriger Arbeiten zur Punktzuordnung

Der Schwerpunkt der vorgelegten Arbeit liegt auf Verfahren zur Punktzuordnung. Hierfür sind bisherige Verfahren nur begrenzt invariant bzw. robust gegen bestimmte häufig auftretende Einflüsse. Zu diesen gehören:

- Änderungen der Lichtverhältnisse
- Globale Maßstabsunterschiede
- Große Blickwinkeländerung
- Wiederkehrende Strukturen

Die Robustheit eines Verfahrens bezüglich einem, mehrerer, oder aller dieser Einflüsse ist je nach Anwendungsgebiet von starker Bedeutung. Der erforderliche Speicher- und Rechenaufwand sind weitere wichtige Aspekte eines Verfahrens.

Änderung der Lichtverhältnisse

Wenn eine Szene zu verschiedenen Zeiten oder mit unterschiedlichen Kameras aufgenommen

wird, führt dies in der Regel dazu, dass die Objekte mit u.U. deutlich verschiedener Intensität abgebildet werden. Die Ursachen dafür können in der Ausleuchtung der Szene, den Reflexionseigenschaften von Objekten und in den Belichtungseinstellungen der Kameras liegen. Eine weitere zeitabhängige Veränderung wird durch Schatten verursacht. Viele gängige Bildzuordnungsverfahren sind nur sehr begrenzt robust gegen stärkere Änderung der Intensität (Mikolajczyk & Schmid, 2005; Mikolajczyk *et al.*, 2005).

Globale Maßstabsunterschiede

Globaler Maßstabsunterschied bedeutet in dieser Arbeit, dass zwei Bilder eine Szene mit einem in allen Bildregionen gleichen oder sehr ähnlichen Maßstabsunterschied abbilden. Dies tritt grundsätzlich bzw. näherungsweise dann auf, wenn die Kamera zwischen den Aufnahmen gerade auf das weiter entfernte Objekt zu oder davon weg bewegt wird, wenn die Kamera gezoomt wird, sowie in der Regel dann, wenn verschiedene Kameras verwendet werden. Weisen Bildpaare einen globalen Maßstabsunterschied auf, dann liefern Bildausschnitt basierte Zuordnungsverfahren nur dann korrekte Ergebnisse, wenn die Ausschnittsgrößen bezüglich des Maßstabsunterschieds angepasst werden. Andernfalls beinhalten Bildausschnitte korrespondierender Punkte im Allgemeinen keine übereinstimmenden Bildregionen, wie dies Abb. 3.8 veranschaulicht.

Punkt-Operatoren, wie Förstner- und Harris-Operator, geben, wenn sie als Ecken-Detektor verwendet werden, wegen der zu Grunde liegenden Methodik grundsätzlich keinen Aufschluss über die Skalierung eines markanten Punktes. Die Ursache dafür liegt in der Ausrichtung der Gradienten in einer Ecken-Region. Diese sind alle tangential zum Zentrum der Ecke ausgerichtet, so dass Bildausschnitte der betroffenen Region unabhängig von der Ausschnittsgröße gleich strukturiert sind. Dies ändert sich erst, wenn umliegende Bildregionen, die nicht zur Ecke gehören, einbezogen werden.

Lindebergs Methode zur maßstabsinvarianten Bildzuordnung (siehe Abschnitt 3.1) ist gut geeignet, um Bildzuordnung trotz globaler Maßstabsunterschiede zu ermöglichen. Allerdings sind Maßstabsraumextremwerte nur in Bildregionen mit Blob-Strukturen gut bestimmbar. Starke Förstner- und Harris-Punkte sind dafür auf Grund der oben geschilderten Problematik ungeeignet. Es wurde ein Ansatz präsentiert, der den Harris Operator um Maßstabsraumextrema erweitert (Mikolajczyk & Schmid, 2004). Allerdings waren die Ergebnisse folgerichtig nicht zufriedenstellend. Förstner *et al.* (2009) präsentierten einen Ansatz, der diesem Problem begegnet, um damit auch für Eckenstrukturen Aufschluss über die Skalierung zu bekommen. Dafür wurden Erkenntnisse aus (Bigün,

Abbildung 3.8: Problematik der bildausschnitt-basierten Verfahren bei globalen Maßstabsunterschieden. Die 13×13 großen Bildausschnitte (unten) um die markanten Punkte (oben, grün) weisen deutlich verschiedene Inhalte auf, obgleich die Bildregionen korrespondieren. Die Rotation um die Hauptachse wurde für die Bildausschnitte angeglichen.

1990) einbezogen. Der aus dieser Arbeit hervorgegangene Punkt-Operator wurde Scale-invariant Feature Operator (SFOP) genannt.

Zusammenfassend ist eine zuverlässige Bildzuordnung bei globalen Maßstabsunterschieden nur mit wenigen bislang vorgestellten Methoden, wie beispielsweise SIFT und SFOP möglich.

Große Blickwinkeländerung

Viele bestehende Anwendungen basieren auf der Auswertung von Videosequenzen, bei denen eine sehr große Anzahl an Bildern vorliegt und geringe Blickwinkeländerungen damit quasi per se gegeben sind. Diese Voraussetzungen sind jedoch nicht für alle Anwendungsfälle erfüllt. So stellen auch Bildverbände aus Einzelbildern u.U. von verschiedenen Kameras mit großer Basis eine wichtige Datenquelle dar (siehe Abschnitt 3.4). Weiterhin verbessern Schleifenschlüsse die Qua-

lität der rekonstruierten 3D Geometrie sehr stark. Hierfür treten auch bei Sequenzen häufig große Blickwinkeländerungen auf. Deshalb werden an dieser Stelle die durch Blickwinkeländerungen auftretenden Schwierigkeiten eingehend diskutiert.

Perspektive Verzerrung tritt bei verschiedenen Blickrichtungen auf eine Szene zwangsläufig auf. Je größer die Blickwinkeländerung zwischen verschiedenen Bildern ist, desto stärker sind die Auswirkungen. Perspektive Verzerrung ist zusätzlich aber auch von der Beschaffenheit der Szene abhängig. Die Auswirkungen von perspektiver Verzerrung sind in Abb. 3.9 veranschaulicht.

Abbildung 3.9: Veranschaulichung der Auswirkung der perspektiven Verzerrung durch Änderung des Blickwinkels. Links: Eine kreisförmig markierte Bildregion. Rechts: Es wurde dieselbe Bildregion markiert, die Blickrichtung auf die Szene ist jedoch eine deutlich andere. Infolgedessen ist die Markierung nicht mehr kreisförmig, sondern elliptisch. Bilder: (Mikolajczyk *et al.*, 2005).

Bei Änderung der Blickrichtung auf eine Szene verändern sich korrespondierende Bildregionen je nach Größe der Änderung deutlich. Für die Bildzuordnung reicht es somit nicht aus, die Lokalisierung eines markanten Punktes genau zu kennen. Während für Translation, Rotation um die Hauptachse und Maßstab leistungsstarke Methoden zur Normalisierung entwickelt wurden, fehlt eine solche bislang für perspektive Verzerrung. Ansätze, die darauf abzielen, liefern vergleichsweise schlechte Ergebnisse hinsichtlich Anzahl und Anteil der korrekten Zuordnungen sowie Genauigkeit (Mikolajczyk & Schmid, 2004).

Der SIFT-Operator zählt mittlerweile zu den am häufigsten verwendeten Punkt-Operatoren. Er ist jedoch nur sehr begrenzt robust gegen perspektive Verzerrung (Lowe, 2004; Morel & Yu, 2009), da er grundsätzlich für Bildsequenzen mit kleiner Basis konzipiert wurde. Die bislang mit Abstand höchste Robustheit gegen perspektive Verzerrung weist ASIFT (siehe Abschnitt 3.3) auf, bei dem affine Transformationen, womit Blickwinkeländerungen simuliert werden, systematisch ausprobiert werden. Obgleich der Simulationsvorgang optimiert wurde, ist ein nicht unerheblicher Rechenaufwand obligatorisch. Ein Gesamtsystem zur 3D Rekonstruktion, welches ASIFT oder das Simulieren von Blickwinkeländerungen allgemein verwendet, wurde noch nicht präsentiert.

Bei großen Blickwinkeländerungen werden korrespondierende Bildregionen in lokal z.T. deutlich verschiedenen Maßstäben abgebildet. Es können Regionen mit höherer, gleicher und niedrigerer Auflösung in einem Bildpaar vorkommen. Eine solche Szene ist in Abb. 3.10 dargestellt. Das Auftreten lokaler Maßstabsunterschiede ist abhängig von der Aufnahmekonfiguration und von der Beschaffenheit der Szene. Anders als bei einem globalen Maßstabsunterschied können auch mit nicht-maßstabsinvarianten Zuordnungsmethoden gute Ergebnisse erzielt werden, wenn eine große Zahl markanter Punkten in Bereichen mit gleichem Maßstab vorhanden sind. Abbildung 3.10 zeigt ein Beispiel für eine 3D Rekonstruktion auf der Grundlage von Förstner-Punkten in Verbindung mit normierter Kreuzkorrelation und affiner Kleinster-Quadrate-Zuordnung.

Grundsätzlich sind Bildpaare mit erheblichen lokalen Maßstabsunterschieden aber nur dann mit hoher Genauigkeit orientierbar, wenn Techniken verwendet werden, die eine korrekte Bildzuordnung bei Maßstabsunterschieden zuverlässig ermöglichen, wofür nur wenige Zuordnungsverfahren in Betracht kommen.

Wiederkehrende Strukturen

Fassaden von Gebäuden weisen häufig wiederkehrende oder ähnliche Strukturen wie z.B. Ziegelsteine auf. Dies führt in Bildpaaren zwangsläufig zu starken Ähnlichkeiten von nicht homologen Regionen und damit zu Fehlzuordnungen. Zur Lösung dieses Problems ist die Einbeziehung der

Abbildung 3.10: Beispiel für eine Aufnahmekonfiguration, bei der erhebliche lokale Maßstabsunterschiede auftreten (oben). Durch die große Blickwinkeländerung werden in den beiden Bildern die Objekte lediglich in einem kleinen Bereich (angedeutet durch rote Markierung) in ungefähr gleicher Auflösung dargestellt. Aufgrund der starken Texturierung ist in diesem Fall eine 3D Rekonstruktion auch auf der Grundlage von grundlegend nicht maßstabsinvarianten Förstner-Punkten möglich (unten).

geometrischen Plausibilität, z.B. Epipolargeometrie oder die Schnitte der Sehstrahlen von drei und mehr Bildern hilfreich.

Neben den am Anfang des Abschnitts genannten, wird im Folgenden noch auf einige weitere oft auftretende Probleme eingegangen.

Oberflächen mit starker Krümmung

Eine Vielzahl von Zuordnungsverfahren verwendet die affine Transformation, um durch Blickwinkeländerungen verursachte perspektive Verzerrung zu korrigieren (Mikolajczyk *et al.*, 2005). Diese

Vorgehensweise beruht u.a. auf der Annahme, dass sich die Objekte der zuzuordnenden Bildregionen jeweils auf einer Ebene befinden. Bei Objekten, deren Oberfläche eine stärkere Krümmung aufweist, ist die Annahme nicht erfüllt. Wird eine affine Transformation für die Transformation des gesamten Bildes verwendet, führt die globale Dehnung abhängig von der Beschaffenheit der Szene zu Fehlern. In allen genannten Fällen wird die Problematik bei größeren Bildausschnitten für die Zuordnung verschärft. Eine affine Transformation ist aber andererseits nur dann sinnvoll als Zuordnungsmethode verwendbar, wenn die verwendeten Bildausschnitte nicht zu klein gewählt werden.

Beschränkte Präzision und Zuverlässigkeit

Obgleich eine Vielzahl durchdachter Methoden zur automatisierten Bildzuordnung verfügbar ist, treten grundsätzlich Fehlzuordnungen auf. Für schwierige Bildpaare ist ein Fehleranteil von mehr als 50% normal. Obgleich mit dem häufig genutzten SIFT-Operator die Möglichkeit besteht, auch bei Maßstabsunterschieden Bildzuordnung durchzuführen, stößt er bei Bildsequenzen mit großer Basis schnell an Grenzen. Zudem ist die Lokalisierung von SIFT-Punkten grundsätzlich ungenauer als beispielsweise von Förstner- oder Harris-Punkten. Um auch mit hohen Fehleranteilen präzise und zuverlässig Bildverbände orientieren zu können, sind weitere Methoden erforderlich, die über die bisherigen Zuordnungsverfahren hinaus gehen.

Hoher Rechen- und Ressourcenaufwand

Viele Punktextraktions- und Bildzuordnungsverfahren erfordern einen hohen Rechen- und Ressourcenaufwand, der die Kapazität von Standard PC Hardware deutlich übersteigen kann. Bei der Bildzuordnung gibt es zwischen zwei Bildern mit m bzw. n markanten Punkten grundsätzlich $m \times n$ mögliche Zuordnungen. Die Anzahl der notwendigen Zuordnungsvorgänge wächst somit quadratisch zur Anzahl der markanten Punkte pro Bild. Wenn für ein Zuordnungsverfahren keine Möglichkeit besteht, den Aufwand zu reduzieren, beispielsweise durch k-d Bäume basierte Sortierverfahren oder die Einbeziehung geometrischer Plausibilität, dann steigt der Zeitaufwand bei Bildern mit Auflösungen im Bereich von 10 Megapixeln oder höher so stark an, dass diese praktisch nicht mehr eingesetzt werden können. Mit ASIFT (siehe Abschnitt 3.3) wurde ein Ansatz präsentiert, der sehr robust gegen viele Störeinflüsse ist, dessen Rechen- und Ressourcenaufwand jedoch beachtlich ist. Obgleich die Simulation der Blickwinkeländerungen hinsichtlich der Laufzeit optimiert wurde, ist der Rechenaufwand mit dem 2,25-fachen der SIFT Zuordnung nach wie vor hoch. Zudem ist der Speicheraufwand von ASIFT bei hoch aufgelösten Bildern sehr hoch.

Kapitel 4

Neuartiger Ansatz zur Bestimmung der Relativen Orientierung

Wie das vorausgegangene Kapitel zeigt, haben derzeitige Methoden zur Punktzuordnung und 3D Rekonstruktion verschiedene Defizite. Sie unterliegen deshalb, je nach Methode, bestimmten Bedingungen. Ziel dieser Arbeit ist es, einen praxistauglichen Ansatz zu entwickeln, dessen Voraussetzungen die Eigenschaften einer Bildmenge möglichst wenig einschränken. Damit soll die Möglichkeit geschaffen werden, auch zufällig zustande gekommene Bildverbände mit großer Basis zu orientieren.

Der neuartige Ansatz FASIAM (Fast Accurate Scale Invariant Affine Matching) verwendet für markante Punkte, die bzgl. ihrer Lokalisierung, Skalierung und Orientierung in der Bildebene normiert sind, eine Zuordnungsmethode, die eine hohe Robustheit gegen möglichst viele Störeinflüsse aufweist. Die Wahl der Methode zur Punktextraktion fiel auf den SIFT-Punkt Operator (siehe Abschnitt 3.1), da dieser nicht nur die genannten Normierungen aufweist, sondern auch durch eine leistungsstarke, leicht integrierbare Softwareimplementierung verfügbar ist (Wu, 2007). Grundsätzlich ist die im Folgenden erläuterte Zuordnungsmethode nicht auf diesen Punkt-Operator begrenzt. Es kann jeder andere verwendet werden, der Information über Lokalisation, Rotation um die Hauptachse und Skalierung bereitstellt. Der SFOP-Operator (Förstner *et al.*, 2009) erscheint für die 3D Rekonstruktion z.B. von Gebäuden besser geeignet, da er mit der Zielsetzung entwickelt wurde, auch in Bildregionen, die für SIFT-Punkte ungeeignet sind, eine hohe Anzahl gut zuzuordnender markanter Punkte zu liefern. Eine leistungsstarke und einfach integrierbare Umsetzung lag zum Zeitpunkt der Fertigstellung dieser Arbeit jedoch nicht vor. Zudem wurde für diesen Operator bislang keine Normierung für die Rotation um die Hauptachse definiert.

Andere Punktoperatoren und Methoden zur Bildzuordnung bei Maßstabsunterschieden lieferten keine zufriedenstellenden Ergebnisse (Mikolajczyk & Schmid, 2004), so dass sie bereits aufgrund dieses Kriteriums für die Zielsetzung dieser Arbeit nicht in Betracht kommen. Im Weiteren wird diskutiert, wie Invarianz bzw. Robustheit gegenüber verschiedenen Störeinflüssen, besonders Maßstabsunterschied und perspektive Verzerrung, erreicht werden kann. Für die SIFT-Punktextraktion wurde hierbei ein modifizierter Parametersatz verwendet, der eine deutlich höhere Anzahl an SIFT-Punkten liefert, als bei der Standardvorgehensweise.

4.1 Robuste Zuordnung bei Änderung der Lichtverhältnisse

Gleichmäßige Lichtverhältnisse können in einem Bildverband nur unter speziellen Umständen vorausgesetzt werden. So müssen die zugehörigen Bilder zur selben Tageszeit aufgenommen werden und bei Verwendung verschiedener Kameras müssen die Belichtungseinstellungen übereinstimmen. Darüber hinaus müssen die Reflexionseigenschaften näherungsweise dem Lambertschen Gesetz entsprechen. Strecha *et al.* (2010) diskutieren die Problematik, dass Bilder einer Szene, die unter deutlich verschiedenen Lichtverhältnissen abgebildet wurde, häufig nicht in einem Gesamtverbund orientiert werden können und präsentieren einen Ansatz basierend auf unabhängigen Teilverbänden von Bildern (siehe Abschnitt 3.4). Ziel der vorgelegten Arbeit ist es jedoch, ein robustes Zuordnungsverfahren umzusetzen, welches diese Problematik in möglichst hohem Maße löst. Die normierte Kreuzkorrelation (siehe Abschnitt 3.2) ist aufgrund der Einbeziehung der Mittelwerte und Varianzen der zugehörigen Bildausschnitte invariant gegen gleichmäßige Veränderung der Lichtverhältnisse. Bei guter Lokalisierbarkeit der markanten Punkte ist diese Methode grundsätzlich gut geeignet für die Bildzuordnung. Die normierte Kreuzkorrelation wird daher im neu entwickelten Ansatz zur Vorauswahl verwendet.

4.2 Invarianz gegen Rotation um die Hauptachse

Punktzuordnung bei Rotation um die Hauptachse ist ohne größere Schwierigkeiten möglich, indem die Rotation bestimmt und berücksichtigt wird. Es wird wie in Abschnitt 3.1 beschrieben vorgegangen: Die vom einzelnen Punkt-Operator bestimmte Orientierung in der Bildebene und die Zuordnung der Punkte werden in Verbindung mit einem Histogramm über die Drehwinkel verwendet, um die relative Rotation zweier Bilder um die Hauptachse robust zu bestimmen. In einem

zweiten Zuordnungsvorgang wird Letztere verwendet, um alle Bildausschnitte entsprechend zu normalisieren. Hierbei ist es grundsätzlich nicht von Bedeutung, welcher Punktoperator verwendet wird. Er muss lediglich die Orientierung in der Bildebene bestimmen können.

4.3 Robuste Zuordnung bei Maßstabsunterschieden

Die Zuordnungsmethode für SIFT-Punkte (siehe Abschnitt 3.1) ist grundsätzlich auch bei Maßstabsunterschieden geeignet. Sie wurde allerdings insbesondere für Bildsequenzen mit kleiner Basis entwickelt. Dementsprechend nimmt die Zahl der korrekten Zuordnungen bei größer werdenden Maßstabsunterschieden schnell ab (siehe z.B. Abb. 6.14). Auch die Erhöhung der Anzahl der Maßstabsraumebenen (siehe Abschnitt 3.1) führt hierbei zu keiner Verbesserung.

Im Folgenden wird untersucht, wie die in Abschnitt 4.1 entworfene Idee, normierte Kreuzkorrelation und damit Bildausschnitt basierte Zuordnung zu verwenden, auf die Zuordnung bei Maßstabsunterschieden erweiterbar ist. Bei Bildausschnitt basierter Zuordnung muss gewährleistet sein, dass korrespondierende Ausschnitte aus zwei verschiedenen Bildern ähnlich rotiert sind und einen ähnlichen Maßstab besitzen, damit eine zuverlässige Zuordnung erreicht wird. Dafür wird (Lindeberg, 1994) zur Bestimmung von Maßstabsraumextremwerten verwendet. Auf der Grundlage des Harris Operators wurde bereits ein ähnlicher Ansatz vorgestellt, bei dem der Strukturtensor (siehe Abschnitt 3.1) entsprechend der Maßstabsraumextremwerte angepasst wurde (Mikolajczyk & Schmid, 2004). Die Resultate waren jedoch nicht zufriedenstellend.

Für die vorgelegte Arbeit wurde die Idee entwickelt, Bilder so zu reskalieren, dass die Inhalte der Bildausschnitte jeweils für zuzuordnende Punkte angepasst werden. Für jeden SIFT-Punkt wird die Maßstabsraumebene mit dem σ bestimmt, auf der der Maßstabsraumextremwert gefunden wurde. Die Kenntnis von σ der zuzuordnenden markanten Punkte ermöglicht die Bestimmung des Skalierungsverhältnisses für die zugehörigen Bildausschnitte.

Die von Lowe (2004) entworfene Methode zur Extraktion von SIFT-Punkten ist für eine deskriptorbasierte Zuordnung optimiert und unterdrückt eine Vielzahl an Punkten, die dafür nicht geeignet sind. Für die Idee der korrelationsbasierten Zuordnung wird auf diese Unterdrückung verzichtet. Dies betrifft die Zahl der Maßstabsraumebenen (Lowe: 3, Ansatz dieser Arbeit: 10), den Toleranzwert für Kantenstrukturen (Lowe: 5, Ansatz dieser Arbeit: 255) und den Schwellwert bei der Bestimmung von Gauß-Ebenen (Lowe: 0,02/3, Ansatz dieser Arbeit: 0,0002/3). Aufgrund der leistungsstarken Softwareimplementierung von Wu (2007) ist eine schnelle Punktextraktion auch mit dieser Parametrisierung möglich.

Der Reskalierungsfaktor r_s ist der Faktor, um den Höhe und Breite herunter skaliert werden müssen, um das zugehörige Bild so zu skalieren, dass es der Maßstabsraumebene entspricht, auf die der Extremwert bestimmt wurde. Dieser ergibt sich aus den im Abschnitt 2.4 erläuterten Eigenschaften der Gaußfunktion zur Glättung von Bildfunktionen und der damit verbundenen Beziehung zur Skalierung von Bildern.
Wegen Maßstabsraumebene L

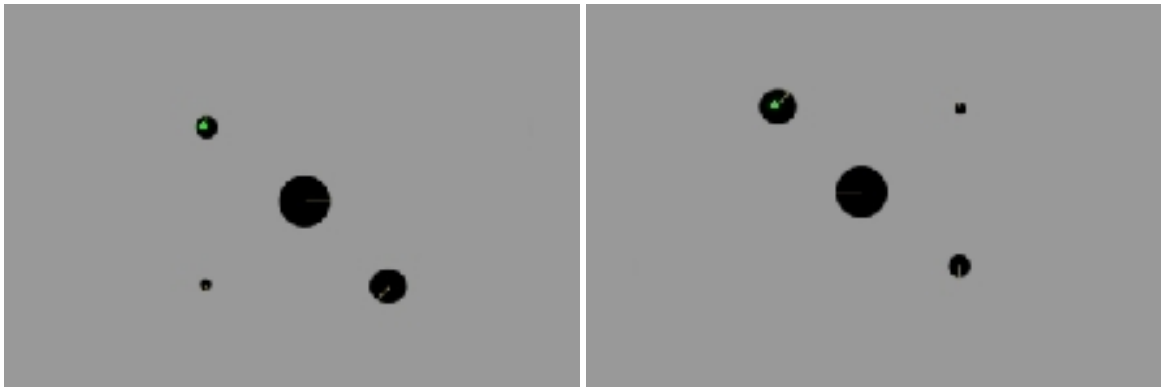
$$L(x,y;\sigma) = \left(\frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} * I \right)(x,y)$$

gilt für den Reskalierungsfaktor r_s:

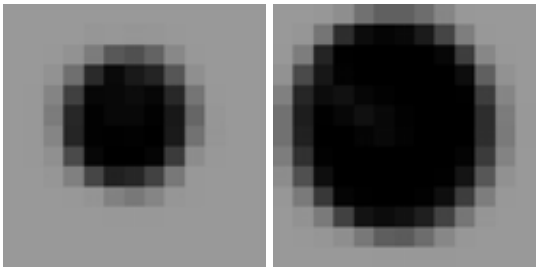
$$r_s = \frac{1}{\sqrt{2}\sigma} \,. \tag{4.1}$$

Um möglichst spezifische Bildausschnitte für die Korrelation zu verwenden, wurde die Größe des Ausschnittsfensters empirisch auf 13×13 Pixel festgelegt. Diese Größe ist ein Kompromiss zwischen Rechenaufwand, Selektivität und Robustheit gegenüber geometrischer Verzerrung. Zur Anpassung der Inhalte der Korrelationsfenster werden die Maßstabsraumextremwerte verwendet (siehe Abb. 4.2). Die jeweils zugehörige Bildregion, die zum Auffinden der Extremwerte verwendet wird, hat eine Größe von 3×3 Pixel (Lindeberg, 1998; Lowe, 2004). Durch das in Gleichung (4.1) beschriebene Verhältnis können damit die Bildauflösungen bestimmt werden, in denen die korrespondierenden Bildregionen durch einen 3×3 Ausschnitt erfasst werden. Sowohl die geringe Größe, als auch die starke Glättung dieser Auflösung macht sie für die Korrelation wenig geeignet, da die zugehörigen Bildausschnitte zu unspezifisch sind. Da das Skalierungsverhältnis über die Maßstabsraumextremwerte bekannt ist, können aber Bildauflösungen und Größe der zugehörigen Bildregionen angepasst werden.

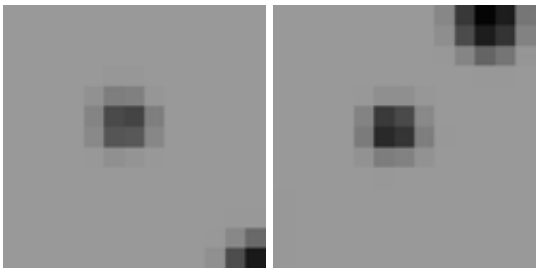
Abbildung 4.1: Links: Vorbereitung eines Bildpaars für die normierte Kreuzkorrelation unter Beachtung der Skalierungsverhältnisse. Rechts: Auswahl der Bildebenen für die normierte Kreuzkorrelation in Abhängigkeit vom Maßstabsraum Extremwert der jeweiligen SIFT-Punkte.



Extrahierte Punkte (grün).



Unskalierte 13×13 Pixel Ausschnitte (Originalauflösung). NCC: 0,64.



13×13 Ausschnitte in den Auflösungen der Maßstabsraumextremwerte. NCC: 0,34.

Abbildung 4.2: Veranschaulichung der Nutzung der Maßstabsraumextremwerte zur Anpassung der Inhalte der Korrelationsfenster (13×13 Pixel). Oben: Zwei synthetische Testbilder mit jeweils einem SIFT-Punkt (grün). Mitte: Zwei Bildausschnitte um die SIFT-Punkte bei Verwendung der Originalbildgröße. Unten: Bildausschnitte bei Verwendung der Skalierungen entsprechend den Maßstabsraumextremwerten (33,7% und 20,3%). Die Ausschnitte sind wiederum um die vom Punkt-Operator bestimmten Drehungen ($10,4^\circ$ und $274,7^\circ$) rotiert.

Für die normierte Kreuzkorrelation (NCC) erscheint es am sinnvollsten, das niedriger aufgelös-

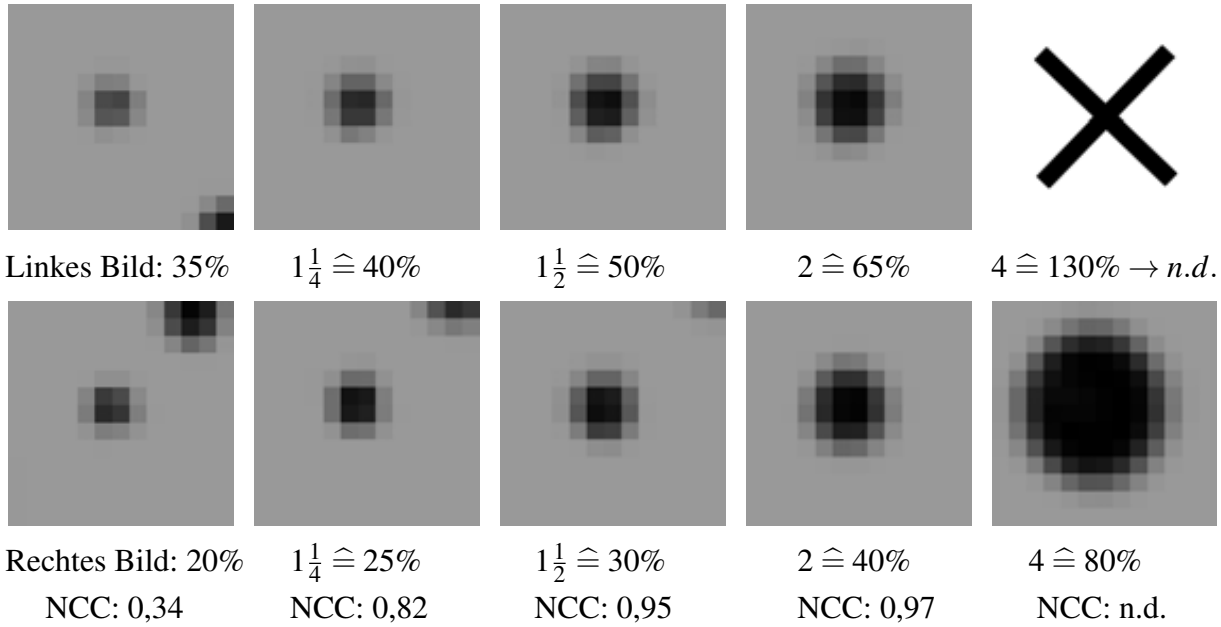


Abbildung 4.3: Fortführung des Beispiels aus Abb. 4.2 – Verwendung der Maßstabsraumextremwerte als Bezugsgröße zur Angleichung der Bildausschnitte. Die Ausschnittsfenster haben jeweils eine Größe von 13×13 Pixel. Die Werte sind auf fünf Prozent gerundet. Da die Bildregion, mit der der Maßstabsraumextremwert bestimmt wird, eine Größe von 3×3 Pixel hat, kann für die gewählte Fenstergröße eine etwas mehr als viermal so hohe Skalierung verwendet werden, sofern aufgrund der Größe des Punktmerkmals nicht die Originalbildgröße überschritten wird. Die Ausschnitte sind um die vom Punkt-Operator bestimmten Drehungen ($10,4°$ und $274,7°$) rotiert.

te Bild in der Originalgröße zu belassen, bzw. möglichst wenig zum Zwecke der Eliminierung sensorbedingten Rauschens zu glätten und das höher aufgelöste so weit herunterzusampeln, dass das erforderliche Skalierungsverhältnis erreicht wird. Für eine performante Software-Implementierung war es jedoch erforderlich, die Anpassungen zu vergröbern, um nicht zu viele Samplings durchführen zu müssen. Dafür werden für jedes Bildpaar zwanzig reduzierte Bilder mit einer Auflösung von 5% bis 95% und 99% der Originalauflösung vorbereitet (siehe Abb. 4.1).

Für jeden SIFT-Punkt werden die Bildausschnitte für insgesamt sechs Ebenen vorbereitet. Dazu gehört die Originalauflösung und die Auflösung, welche dem Maßstabsraumextremwert zugehörig ist. Die vier übrigen liegen jeweils auf einer Ebene, welche um einen bestimmten Faktor höher aufgelöst ist, als die dem Maßstabsraumextremwert zugehörige Auflösung. Die Faktoren lauten $1\frac{1}{4}$, $1\frac{1}{2}$, 2 und 4. Das zugrundeliegende Konzept ist in Abb. 4.3 veranschaulicht. Der Maßstabsraum-

extremwert und die zugehörige Bildauflösung werden als Bezugsgröße verwendet, um die Inhalte der Bildausschnitte anzugleichen. Dabei wird ausgenutzt, dass die Bildausschnitte bei korrespondierenden Bildregionen auf der Maßstabsraumextremwert-Ebene angepasste Inhalte aufweisen.

Abbildung 4.4: Anpassung der Bildausschnitte zur Lösung der in Abschnitt 3.5 erläuterten Problematik. Obere Zeile: Bilder mit markierten SIFT-Punkten. Mittlere Zeile: 13×13 große Bildausschnitte der Originalauflösung, die offensichtlich verschiedene Inhalte aufweisen. Untere Zeile: 13×13 große Bildausschnitte, die mit der in diesem Abschnitt beschriebenen Methode angepasst wurden. Die Inhalte stimmen nahezu überein.

Die Reskalierungsfaktoren werden auf die vorbereiteten Auflösungen (siehe Abb. 4.1) gerundet.

Der daraus resultierende Fehler beim Skalierungsverhältnis wird in Kauf genommen. Für jede Ebene des Bildausschnitts werden Mittelwert und Varianz berechnet und gespeichert, um eine performante Berechnung der Kreuzkorrelation zu ermöglichen. Der Rechen- und Speicheraufwand ist trotz der beschriebenen Vergröberung beachtlich, aber akzeptabel.

Ein Beispiel für reale Bilder ist in Abb. 4.4 veranschaulicht. Ein $\sigma_1 = 2,67$ im linken und $\sigma_2 = 1,73$ der zugehörigen SIFT-Punkte im rechten Bild führen bei Anwendung von Gleichung (4.1) zu den Reskalierungsfaktoren $r_{s1} = 0,26$ und $r_{s2} = 0,41$. Das bedeutet, dass der bestimmte Maßstabsraumextremwert bei 26% bzw. bei 41% der Originalauflösung liegt. Aus dieser Information folgt das Skalierungsverhältnis der zugehörigen Bildausschnitte. Der linke Bildausschnitt ist um den Faktor $1,58$ höher aufgelöst, als der rechte.

Die Bildausschnitte weisen dieselben Inhalte auf, solange dieses Skalierungsverhältnis bewahrt bleibt. Werden also die Auflösungen, auf denen die Maßstabsraumextremwerte bestimmt wurden, jeweils um denselben Faktor erhöht, so beinhalten die Bildausschnitte nach wie vor identische Bildregionen, aber in einer höheren Auflösung. Bei diesem Beispiel wurde der Faktor 2 verwendet, so dass der linke Ausschnitt 52% und der rechte 82%, respektive nach Rundung auf 5%, 50% und 80% der Originalauflösung aufweist.

4.4 Robuste Bildzuordnung bei starken Blickwinkeländerungen

Bei mehr oder weniger zufällig zustande gekommenen Bildmengen können keine kleinen Blickwinkeländerungen vorausgesetzt werden. Somit können zwischen den Bildausschnitten substantielle perspektive Verzerrungen auftreten (siehe Abschnitt 3.5). Um diesem Problem zu begegnen, bestehen im Wesentlichen folgende Ansätze:

1. Normierung eines Bildausschnitts bezüglich der affinen Transformationsparameter (Mikolajczyk & Schmid, 2004).

2. Lokale Optimierung durch Affine Kleinste-Quadrate-Zuordnung.

3. Simulation von Blickpunktänderung durch affine Transformation (ASIFT) (Morel & Yu, 2009).

Alle Ansätze verwenden jeweils die affine Transformation und setzen somit voraus, dass die zuzuordnenden Bildregionen auf einer Ebene liegen. Perspektive Verzerrung kann damit nur näherungsweise und lokal approximiert werden. Die von Mikolajczyk & Schmid (2004) präsentierten Ansätze Harris- und Hessian-Affine liefern keine zufriedenstellenden Resultate. Über die Ursachen wird u.a. in (Morel & Yu, 2009) diskutiert und hierbei festgestellt, dass im Wesentlichen nicht beachtet wurde, dass das Hochsampeln von Bildregionen im Allgemeinen zu keinen sinnvollen Ergebnissen führt (siehe Abschnitt 2.4). Abhängig von der Beschaffenheit der Bildregionen erfolgt dies aber bei der Angleichung der perspektiven Verzerrung. Der Ansatz von (Morel & Yu, 2009) Blickwinkeländerungen zu simulieren ist neu und wurde in der Entstehungsphase dieser Arbeit veröffentlicht.

Die Affine Kleinste-Quadrate-Zuordnung wendet auf zwei zuzuordnende Bildausschnitte die affine Transformation an und optimiert die Transformationsparameter. Des Weiteren kann auf diese Weise die Lokalisierung der korrespondierenden Bildpunkte nicht nur subpixelgenau bestimmt werden, sondern auch noch die Zuordnungsgenauigkeit in Form von Varianz und Kovarianz bestimmt werden. Da der Rechenaufwand der affinen Kleinsten-Quadrate-Zuordnung geringer ist als das aufwändige Austesten von Blickwinkeländerungen, fiel die Wahl auf diese Methode. Weil Bildausschnitte verwendet werden, besteht wie beim Korrelieren die in Abschnitt 3.5 beschriebene Problematik, dass bei Maßstabsunterschieden sichergestellt sein muss, dass die Ausschnitte entsprechend der Auflösung angeglichen werden. Aus diesem Grund wird analog zur im vorangegangenen Abschnitt 4.3 beschriebenen Methode vorgegangen. Für die Affine Kleinste Quadrate Zuordnung konnte dieses Konzept jedoch einfacher umgesetzt werden. Die Anzahl der durchzuführenden Zuordnungen sind durch die Vorauswahl der Kreuzkorrelation deutlich geringer, so dass die Skalierungen der Bildausschnitte für jedes Punktepaar individuell angepasst werden können. Dementsprechend werden die Bilder nur so weit skaliert, dass die Inhalte der jeweiligen Bildausschnitte angeglichen sind.

4.5 Robuste Zuordnung bei wiederkehrenden Strukturen

Von Menschen geschaffene Strukturen weisen gewöhnlich Symmetrien und Systematiken auf. Fassaden zeigen häufig wiederkehrende Elemente wie Fenster, Ziegelsteine, Fugen oder Verkleidungen. Folglich ähneln sich in Bildern von Fassaden viele Bildregionen sehr stark, was zu vielen Fehlzuordnungen führt. Durch die in Abschnitt 2.3 beschriebene Einbeziehung der geometrischen Plausibilität kann dieses Problem zum größten Teil effizient gelöst werden. Die relative

Orientierung eines Bildpaares kann wie in Abschnitt 2.1 beschrieben durchgeführt werden. Für die Punktextraktion und Bildzuordnung werden die in den vorangegangenen Abschnitten erläuterten Methoden verwendet. Mittels SIFT-Punkten, normierter Kreuzkorrelation sowie affiner Kleinster-Quadrate-Zuordnung unter Verwendung skalierter Bildausschnitte werden Punktkorrespondenzen bestimmt. Aus Letzteren wird robust mittels RANSAC und GRIC (siehe Abschnitt 2.3) die zugehörige Epipolargeometrie bestimmt, mit der die Zuordnungen überprüft werden können. Durch die Überprüfung der geometrischen Plausibilität inklusive Verbesserung mittels Bündelausgleichung kann auch bei einem hohen Anteil an Fehlzuordnungen die korrekte Lösung gefunden werden (Bartelsen & Mayer, 2010a).

4.6 Robuste Zuordnung für Bildtripel

Die Einbeziehung der geometrischen Plausibilität ist für drei Bilder besonders wirkungsvoll. Anders als beim Bildpaar, bei dem es durch die Abbildung eines Punktes ins andere Bild als Epipolarlinie zu Mehrdeutigkeiten kommt, wird beim Bildtripel im Allgemeinfall eine eindeutige Zuordnung erreicht. Daher ist für verlässliche Resultate die Verwendung von Bildtripeln in Verbindung mit robuster Parameterschätzung von (im vorliegenden Fall kalibrierten) Trifokaltensoren sinnvoll (Torr & Zisserman, 1997). Konkret wird beim vorgestellten Ansatz ein Bild als Hauptbild ausgewählt. Zwischen dem Haupt- und den zwei Nebenbildern wird zunächst die Epipolargeometrie und daraus der Trifokaltensor bestimmt.

Mit dem Trifokaltensor ist es möglich, aus Punktkorrespondenzen in zwei Bildern den korrespondierenden Punkt im dritten Bild zu prädizieren (siehe Abb. 4.5). Dazu wird ein Punkt im ersten Nebenbild als Epipolarlinie in das Hauptbild abgebildet. In der Praxis kommt es dabei in der Regel zu Mehrdeutigkeiten, d.h. mehrere Punkte, die von der Bildzuordnung verwendet wurden, liegen auf dieser Linie. Durch die Verwendung des Trifokaltensors kann mittels einer von der Epipolarlinie verschiedenen Linie, die aber den korrespondierenden Punkt enthält, eine Homographie zwischen den Nebenbildern bestimmt werden. Typischerweise wird für die Linie im Hauptbild die Senkrechte der Epipolarlinie verwendet (Mayer, 2002). Die Mehrdeutigkeit des Zweibildfalls wird durch die Homographien aufgelöst. Verschiedene Punkte auf der Epipolarlinie führen zu unterschiedlichen Homographien vom ersten auf das zweite Nebenbild. Diese Vorgehensweise wurde der expliziten Bestimmung der 3D Punkte vorgezogen, da sie hinsichtlich der Laufzeit effizienter ist. Der Schnitt von Epipolarlinien im Hauptbild, der bei Parallelität, die bei Seitwärtsbewegung

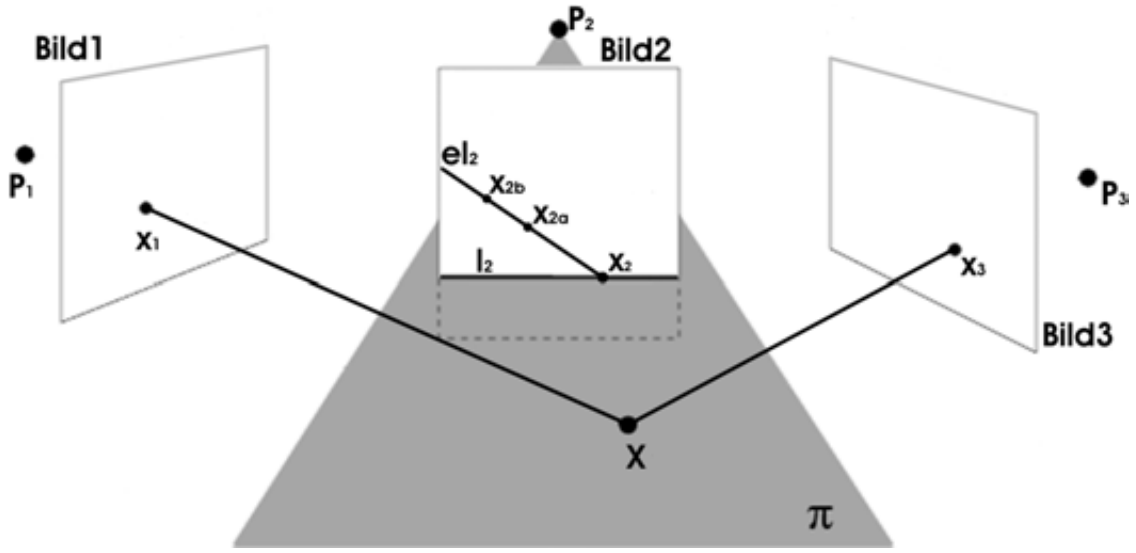


Abbildung 4.5: Verwendung des Trifokaltensors zur Prädiktion von Punkten durch Bestimmung der Homographie zwischen zwei Bildern. Der Punkt x_1 wird als Epipolarlinie el_2 ins Bild 2 abgebildet. Eine von der Epipolarlinie verschiedene Linie l_2, die aber den Punkt x_2 enthält, kann dazu verwendet werden, eine Homographie $\mathbf{H}$ zwischen Bild 1 und Bild 3 mit der Eigenschaft $x_3 = \mathbf{H}x_1$ zu bestimmen. Die Punkte x_{2a} und x_{2b} würden dementsprechend zu anderen Punkten x_{3a} und x_{3b} führen. Skizze: Hartley & Zisserman (2004), Beschriftung und mehrere Details geändert und ergänzt.

häufig genähert auftritt, undefiniert ist, wird bei Verwendung des Trifokaltensors vermieden und man erhält auch für diesen für Bodenaufnahmen typischen Fall eine eindeutige Lösung.

4.7 Kombination der robusten Zuordnungsmethoden

Die in den vorangegangen Abschnitten erläuterten Methoden werden zu einem Ansatz zur robusten Bildzuordnung kombiniert. Grundlage ist die normierte Kreuzkorrelation und die Affine Kleinste-Quadrate-Zuordnung unter Einbeziehung der Skalierungsverhältnisse, wie dies in Abschnitt 4.3 beschrieben ist. Der Ansatz umfasst folgende Schritte:

1. Generierung von Bildpyramiden

2. Bestimmung der Rotation um die Hauptachse (NCC, Bildpaar, niedrige Bildauflösung, Histogramm)
3. Bestimmung von Punktkorrespondenzen unter Berücksichtigung der Rotation um die Hauptachse (NCC, Bildpaar, niedrige Bildauflösung, mehr Punkte)
4. Affine Kleinste-Quadrate-Zuordnung (Bildpaar, niedrige Bildauflösung)
5. Robuste Parameterschätzung (Essentielle Matrix)
6. Bestimmung von Punktkorrespondenzen unter Einbeziehung der geometrischen Plausibilität durch Epipolarlinien (höhere Bildauflösung)
7. Affine Kleinste-Quadrate-Zuordnung (Bildtripel, höhere Bildauflösung)
8. Robuste Parameterschätzung (kalibrierter Trifokaltensor)
9. Die Schritte 6-8 können optional mit noch höherer Bildauflösung wiederholt werden. Hierbei wird der Trifokaltensor zur Prädiktion von Punkten im dritten Bild verwendet, nachdem Zuordnungen in zwei Bildern vorliegen.

4.8 Verknüpfung der Tripel

Auf der Grundlage von Bildtripeln, die jeweils in einem lokalen dreidimensionalen euklidischen Koordinatensystem orientiert sind, kann eine Bildmenge von n Bildern, was unter Annahme einer linearen Sequenz $(n-2)$ Bildtripeln entspricht, unter bestimmten Bedingungen in einen Gesamtverbund überführt werden. Die Sequenz entspricht dem einfachsten Fall. In diesem Abschnitt wird davon ausgegangen, dass die Bildtripel in der Sequenz so sortiert sind, dass die letzten beiden Kamerapositionen von Tripel i mit den ersten beiden von Tripel $i+1$ übereinstimmen. Eine umkehrbar eindeutige Koordinatentransformation zwischen zwei dreidimensionalen euklidischen Vektorräumen kann durch eine 3D Ähnlichkeitstransformation bestehend aus Rotation $\boldsymbol{R}$, Translation t und Maßstab m realisiert werden (siehe Abschnitt 2.2). Basierend darauf ist die Überführung zweier relativ orientierter Bildtripel in ein einheitliches Koordinatensystem grundsätzlich dann möglich, wenn die Projektionsmatrizen bekannt sind und zwei Kameraposition in beiden Tripeln enthalten sind. Es seien $\boldsymbol{P}_{A0}$, $\boldsymbol{P}_{A1}$ und $\boldsymbol{P}_{A2}$ sowie $\boldsymbol{P}_{B0}$, $\boldsymbol{P}_{B1}$ und $\boldsymbol{P}_{B2}$ die 3×4 Projektionsmatrizen zweier

relativ orientierter Bildtripel. Bei Übereinstimmung der Kamerapositionen $\boldsymbol{P}_{A1}$ und $\boldsymbol{P}_{A2}$ mit $\boldsymbol{P}_{B0}$ und $\boldsymbol{P}_{B1}$ mit

$$\boldsymbol{P}_{Ai} = \boldsymbol{R}_{Ai}[\boldsymbol{I}|t_{Ai}] \text{ und } \boldsymbol{P}_{Bi} = \boldsymbol{R}_{Bi}[\boldsymbol{I}|t_{Bi}]$$

gilt wegen Korrespondenz von $\boldsymbol{P}_{A1}$ und $\boldsymbol{P}_{B0}$ für die Rotationsmatrix $\boldsymbol{R}$:

$$\boldsymbol{R}_{B0}\boldsymbol{R} = \boldsymbol{R}_{A1} \;\Rightarrow\; \boldsymbol{R} = \boldsymbol{R}_{B0}^{-1}\boldsymbol{R}_{A1}.$$

Für den Maßstab m ergibt sich wegen der Korrespondenz von $\boldsymbol{P}_{A1}, \boldsymbol{P}_{A2}$ mit $\boldsymbol{P}_{B0}, \boldsymbol{P}_{B1}$:

$$m = \frac{|t_{A2} - t_{A1}|}{|t_{B1} - t_{B0}|}.$$

Für die Translation t gilt:

$$t = m\boldsymbol{R}t_{B0} - t_{A1}$$

Folglich kann die Projektionsmatrix $\boldsymbol{P}_{B2}$ in das Koordinatensystem des Bildtripels A als $\boldsymbol{P}_{A3}$ wie folgt durch die 3D Ähnlichkeitstransformation bestehend aus $\boldsymbol{R}, m$ und t. transformiert werden:

$$\boldsymbol{P}_{A3} = [m * \boldsymbol{R} * \boldsymbol{R}_{B2} | m * \boldsymbol{R} * t_{B2} - t]$$

Die Transformation der 3D-Punkte erfolgt entsprechend. Auf diese Weise können einem Bildtripel beliebig viele Bilder hinzugefügt werden, so dass die relative Orientierung des ersten Bildtripels über die Gesamtsequenz propagiert wird. Diese Vorgehensweise ist nicht auf Bildtripel beschränkt. Auch größere, relativ orientierte Teilmengen können auf dieselbe Weise in ein einheitliches Koordinatensystem überführt werden, einzige Voraussetzung ist die Übereinstimmung zweier Kamerapositionen.

In dieser Arbeit wird bei dieser sequentiellen Vorgehensweise nach jedem Hinzufügen eines Tripels eine Bündelausgleichung (siehe Abschnitt 2.3) durchgeführt, um einen stärkeren Drift der Sequenz zu vermeiden. Bei $(n-2)$ Bildtripeln müssen dementsprechend $(n-3)$ Bündelausgleichungen über insgesamt

$$\sum_{i=1}^{n} i - \sum_{i=1}^{3} i \;=\; \left(\frac{n*(n+1)}{2} - \frac{3*4}{2}\right)$$

Kamerapositionen bestimmt werden.

Um die Anzahl und die Komplexität der Bündelausgleichungen über große Bildmengen und den damit verbundenen hohen Rechenaufwand zu reduzieren, ist es sinnvoll, eine Bildmenge nicht bildweise, sondern über Teilmengen zu verknüpfen. Wenn man erneut davon ausgeht, dass aus einer Bildmenge jeweils lokal orientierte Bildtripel gebildet worden sind, können diese hierarchisch

verknüpft werden. Dazu werden im ersten Schritt Bildtripel zu (relativ orientierten) Teilmengen von vier Bildern verknüpft, was zu $(n-2)/2$ Bündelausgleichungen über jeweils vier Bilder führt. Im zweiten Schritt werden die Teilmengen des ersten Schrittes zu Teilmengen von sechs Bildern verknüpft. Diese Vorgehensweise wird wiederholt, bis die komplette Bildmenge relativ orientiert ist. Die Anzahl der notwendigen Iterationsschritte i beträgt folglich $i = ld(n-2)$ und die Anzahl der erforderlichen Bündelausgleichungen somit

$$\sum_{j=1}^{i} \frac{(n-2)}{2^j}$$

über insgesamt

$$\sum_{j=1}^{i} (2+2^j) * \frac{(n-2)}{2^j}$$

Kamerapositionen. Das bedeutet, dass bei einer Menge von sechs Bildern für den sequentiellen Fall drei Bündelausgleichungen über insgesamt 15 Kamerapositionen, im hierarchischen Fall drei Bündelausgleichungen über 14 Kamerapositionen durchgeführt werden müssen. Bei zehn Bildern müssen im sequentiellen Fall sieben Bündelausgleichungen über insgesamt 49 Kamerapositionen, im hierarchischen Fall sieben Bündelausgleichungen über 38 Kamerapositionen durchgeführt werden. Bereits bei Bildmengen von sechs Bildern ist somit der Rechenaufwand für die hierarchische Verknüpfung geringer. Die in diesem Abschnitt erläuterten Verknüpfungsmethoden sind direkt für Aufnahmekonfigurationen geeignet, die gezielt zur 3D Rekonstruktion eines Objekts erstellt wurden. Bei zufällig zu Stande gekommenen Bildverbänden kann eine linienförmige sequentielle Verknüpfungsreihenfolge der Bildtripel nicht vorausgesetzt werden. Für Aufnahmekonfigurationen wie in Abb. 4.6 ist die mehrfache Verknüpfung eines Bildes mit verschiedenen Tripeln oder Teilsequenzen notwendig.

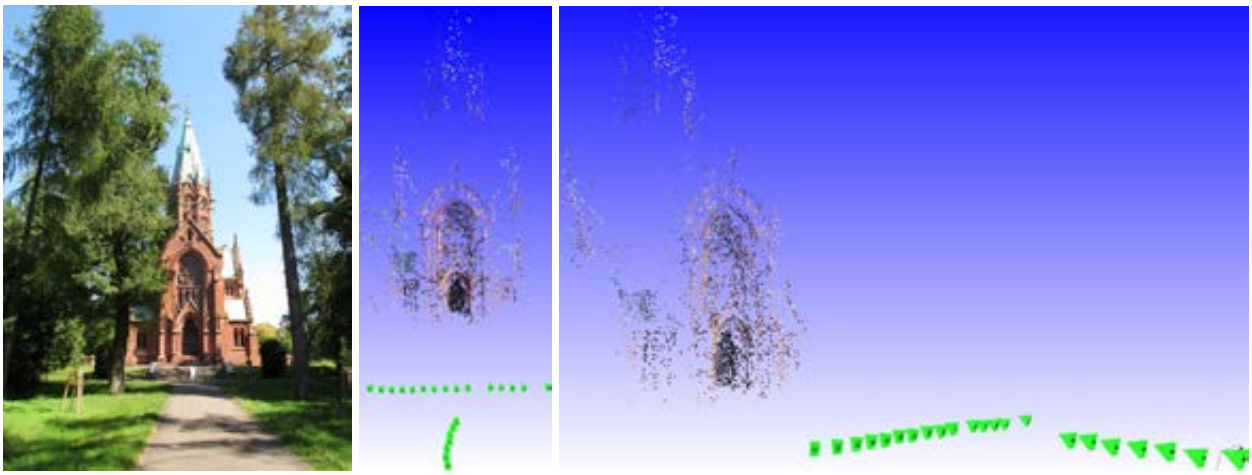


Abbildung 4.6: Beispiel für eine Aufnahmekonfiguration, die aufgrund der T-Struktur nicht auf eine linienförmige Sequenz beschränkt werden kann. Links: Bild aus der verwendeten Bildmenge. Mitte und Rechts: Relative Orientierung – Großherzogliche Grabkapelle Karlsruhe (22 Bilder).

Kapitel 5

Bestimmung der Absoluten Orientierung

Kapitel 3 und 4 zeigen, dass die relative Orientierung einer Bildmenge ohne Verwendung von Näherungen oder Passpunkten mit hoher Genauigkeit bestimmt werden kann. Für die absolute Orientierung ist der Bezug zum Weltkoordinatensystem erforderlich. Um nicht auf Passpunkte angewiesen zu sein, kann bei bekanntem relativen Modell der Szene eine vorhandene Zuordnung der Kamerapositionen in Modell- und Weltkoordinaten genutzt werden. In den letzten Jahren sind kostengünstige Kameras mit integriertem GPS-Empfänger auf den Markt gekommen. Aus diesem Grund wird in diesem Kapitel erläutert, wie die Absolute Orientierung auf der Grundlage von Bildern einer GPS Kamera und deren genauen Relativen Orientierung bestimmt werden kann.

5.1 Schätzung der 3D Ähnlichkeitstransformation

Eine 3D Ähnlichkeitstransformation besteht aus Translation, 3D Rotation und Maßstab (siehe Abschnitt 2.2). Eine solche Abbildung zwischen zwei dreidimensionalen euklidischen Vektorräumen wird durch drei Punktkorrespondenzen eindeutig definiert. Die einzige notwendige und hinreichende Bedingung dabei ist, dass die Punkte nicht kollinear sein dürfen (Schödlbauer, 1995). Jedes nicht kollineare Tripel, bestehend aus korrespondierenden Kamerapositionen in Welt- und Modellkoordinatensystem definiert somit die umkehrbar eindeutige Abbildung f. Aufgrund von Messungenauigkeiten der Weltkoordinaten, z.B. durch eine GPS Kamera, ist die Abbildung jedoch nur als Näherung zu betrachten.

Für die Abbildung von Modell- auf Weltkoordinaten wurde in dieser Arbeit folgende Funktion $f :: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ definiert:

- Beobachtungen: Kamerapositionen im Universal Transversal Mercator (UTM) Weltkoordinatensystem (E – Rechtswert, N – Hochwert und h – Höhe) und zugehörige Positionen im lokalen 3D Modell (x, y und z)

- 7 Unbekannte: Rotation (3 Parameter in der Rotationsmatrix $\boldsymbol{R}$ oder Elemente des normierten Quaternions a, b, c und d), Maßstab (1 Parameter m), und Translation (3 Parameter t_x, t_y und t_z)

$$f :: \begin{pmatrix} E \\ N \\ h \end{pmatrix} = m \cdot \boldsymbol{R} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{5.1}$$

mit

$$\boldsymbol{R} = \begin{pmatrix} 1-2(c^2+d^2) & 2(bc-ad) & 2(bd+ac) \\ 2(bc+ad) & 1-2(b^2+d^2) & 2(cd-ab) \\ 2(bd-ac) & 2(cd+ab) & 1-2(b^2+c^2) \end{pmatrix}$$

Für drei zwischen Welt- und Modellkoordinaten korrespondierende Punktepaaren P_1, P_2, P_3 und p_1, p_2, p_3, mit $P_1 = (E_1, N_1, h_1)^T$ und $p_1 = (x_1, y_1, z_1)^T$ (P_2, P_3 und p_2, p_3 analog) werden die Näherung für Translation, Maßstab und Rotation folgendermaßen bestimmt: Zunächst wird aus numerischen Gründen der Ursprung des Modellkoordinatensystems in den Schwerpunkt aller n-Kamerapositionen verschoben, mit dem Schwerpunkt

$$s_p = \frac{1}{n} \cdot (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} z_i)^T.$$

Analog wird der Schwerpunkt S_P der Kamerapositionen im Weltkoordinatensystem bestimmt, welcher dann der Translation $(t_x, t_y, t_z)^T$ entspricht, da nach der Verschiebung des Modellkoordinatensystems gilt: $s_p = (0,0,0)^T$.
Der Maßstab m folgt aus dem Verhältnis der euklidischen Abstände zweier korrespondierender Punktepaare. Da für eine Näherung drei korrespondierende Punktepaare benötigt werden, werden für die drei möglichen Kombinationen die Maßstäbe m_1, m_2 und m_3 bestimmt. Für die Abbildung von Modell- auf Weltkoordinaten gilt entsprechend:

$$m_1 = \frac{|P_1 - P_2|}{|p_1 - p_2|},\ m_2 = \frac{|P_1 - P_3|}{|p_1 - p_3|} \text{ und } m_3 = \frac{|P_2 - P_3|}{|p_2 - p_3|}.$$

Als Näherung für den Maßstab m wird der Mittelwert dieser drei Werte verwendet.
Zur Bestimmung einer Näherung für die Rotationsmatrix $\boldsymbol{R}$ wird für die drei korrespondierenden

Punktepaare jeweils im Welt- und im Modellkoordinatensystem eine orthonormale Basis bestimmt. Diese hat genau dann den Rang drei, wenn die Basisvektoren der Punkte linear unabhängig sind. Andernfalls ist eine eindeutige Bestimmung der Rotation nicht möglich. Dies entspricht der Bedingung, dass die für eine Näherung verwendeten Punkte in ihrem Koordinatensystem nicht kollinear sein dürfen.
Die Bestimmung der orthonormalen Basis für die Weltkoordinaten folgt aus:

$$U = \frac{P_2 - P_1}{|P_2 - P_1|},\ W = \frac{U \times (P_3 - P_1)}{|U \times (P_3 - P_1)|} \text{ und } V = W \times U$$

Die Bestimmung der orthonormalen Basis u, v und w für die Modellkoordinaten erfolgt analog. Aus den Basisvektoren U, V, W und u, v, w folgen die Rotationsmatrizen $\boldsymbol{R}_1$ und $\boldsymbol{R}_2$ mit

$$\boldsymbol{R}_1 = \begin{pmatrix} U_E & V_E & W_E \\ U_N & V_N & W_N \\ U_h & V_h & W_h \end{pmatrix} \text{ und } \boldsymbol{R}_2 = \begin{pmatrix} u_x & v_x & w_x \\ u_y & v_y & w_y \\ u_z & v_z & w_z \end{pmatrix}.$$

Aufgrund der Punktkorrespondenzen bilden $\boldsymbol{R}_1$ und $\boldsymbol{R}_2$ unter Vernachlässigung von Translation und Maßstab jeweils aus ihrem Urbildraum auf den identischen Bildraum ab. Für die Abbildung von Modellkoordinaten (x, y, z) gilt damit:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \boldsymbol{R}_1^{-1} \boldsymbol{R}_2 \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Somit kann die Rotationsmatrix $\boldsymbol{R}$ mittels

$$\boldsymbol{R} = \boldsymbol{R}_1^T \boldsymbol{R}_2$$

bestimmt werden. Für die Repräsentation einer Rotationsmatrix wird in dieser Arbeit für eine eindeutige Beschreibung ein normiertes Quaternion verwendet.

5.2 Konkretes Funktionalmodell der vermittelnden Ausgleichung

Es wird folgendes Ausgleichungsprinzip für die Unbekannten $\hat{x}$ verwendet:

$$\hat{x} = (\boldsymbol{A}^T \boldsymbol{P} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{P} b$$

Hierbei steht $\boldsymbol{A}$ für die Designmatrix, $\boldsymbol{P}$ die Gewichtsmatrix und b für die Beobachtungen.

Bei der Ableitung der Designmatrix $\boldsymbol{A}$ kann zunächst aufgrund der Überbestimmung des normierten Quaternions und der Eigenschaft $1 = a^2 + b^2 + c^2 + d^2$ eine Unbekannte der Rotationsmatrix substituiert werden, so dass insgesamt sieben Unbekannte verbleiben. Aus numerischen Gründen wird stets das Element mit dem höchsten Betrag substituiert. Im Weiteren werden die drei Unbekannten des Quaternions mit q_1, q_2 und q_3 bezeichnet.
Durch Ausmultiplizieren von Gleichung (5.1) kann der Wertebereich von f zeilenweise bestimmt werden:

$$f_E = m \cdot ((1 - 2(c^2 + d^2)) \cdot x + 2(bc - ad) \cdot y + 2(bd + ac) \cdot z) + t_x \tag{5.2}$$

$$f_N = m \cdot (2(bc + ad) \cdot x + (1 - 2(d^2 + b^2)) \cdot y + 2(cd - ab) \cdot z) + t_y \tag{5.3}$$

$$f_h = m \cdot (2(bd - ac) \cdot x + 2(cd + ab) \cdot y + (1 - 2(b^2 + c^2)) \cdot z) + t_z \tag{5.4}$$

Die Abbildung einer Modellkoordinate $(x, y, z)^T$ auf ein Element von f_E, f_N oder f_h entspricht damit einer Beobachtung. Die Differenz zwischen den so bestimmten und den gemessenen Weltkoordinaten entspricht den abgeleiteten Beobachtungen b. Die $(n \times 7)$ Designmatrix $\boldsymbol{A}$ besteht aus den partiellen Ableitungen von f über die sieben Unbekannten:

$$\boldsymbol{A} \doteq \begin{pmatrix} \frac{\partial f_{E1}}{\partial m} & \frac{\partial f_{E1}}{\partial q_1} & \frac{\partial f_{E1}}{\partial q_2} & \frac{\partial f_{E1}}{\partial q_3} & 1 & 0 & 0 \\ \frac{\partial f_{N1}}{\partial m} & \frac{\partial f_{N1}}{\partial q_1} & \frac{\partial f_{N1}}{\partial q_2} & \frac{\partial f_{N1}}{\partial q_3} & 0 & 1 & 0 \\ \frac{\partial f_{h1}}{\partial m} & \frac{\partial f_{h1}}{\partial q_1} & \frac{\partial f_{h1}}{\partial q_2} & \frac{\partial f_{h1}}{\partial q_3} & 0 & 0 & 1 \\ \frac{\partial f_{E2}}{\partial m} & \frac{\partial f_{E2}}{\partial q_1} & \frac{\partial f_{E2}}{\partial q_2} & \frac{\partial f_{E2}}{\partial q_3} & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Die Anzahl der Beobachtungen n ist abhängig von der Anzahl der Kamerapositionen, für die eine GPS Position gemessen wurde. Unter Berücksichtigung von Gleichungen (5.2), (5.3) und (5.4) gilt somit:

$$n = 3 \cdot (\text{Anzahl Kamerapositionen mit GPS Information})$$

5.3 Generierung geeigneter Näherungen

Da Näherungen für die 3D Ähnlichkeitstransformation durch grobe Fehlmessungen ebenfalls grob falsch sein können, ist ein Kriterium zur Unterscheidung von guten und schlechten Näherungen

erforderlich. Es wurde eine Methode entwickelt, um die geometrische Plausibilität einer Näherung für eine 3D Ähnlichkeitstransformation zwischen Welt- und Modellkoordinaten zu bewerten. Bisher wurden ausschließlich Bildsequenzen mit einer verhältnismäßig geringen Anzahl (ca. 30-40) von Bildern einer GPS Kamera verwendet. Hierfür können aus n GPS-Kamerapositionen alle $\binom{n}{3}$ möglichen Näherungen bestimmt und überprüft werden. Sofern eine deutlich höhere Anzahl an GPS-Kamerapositionen zur Verfügung steht, ist eine zufallsbasierte Vorgehensweise, wie z.B. RANSAC (siehe Abschnitt 2.3), sinnvoll. Grundsätzlich orientiert sich die vorgestellte Arbeit an dem Gedanken des Konsenses (Consensus) in Fischler & Bolles (1981). Für jede der $\binom{n}{3}$ Näherungen werden alle Kamerapositionen des Modellkoordinatensystems, denen eine Weltkoordinate zugeordnet ist, auf das Weltkoordinatensystem abgebildet. Diese berechneten Positionen werden dann mit den zugeordneten verglichen. Für die Bestimmung eines geeigneten Kriteriums für Inlier werden die empirischen Ergebnisse aus Abschnitt 6.3 herangezogen, die zeigen, welche Genauigkeit der Positionsbestimmung für einer GPS Kamera zu erwarten ist.

Kapitel 6

Experimente

Zur Einschätzung der Robustheit des neuen Ansatzes gegenüber häufig auftretenden Störeinflüssen wurde dieser auf der Grundlage des quasi Benchmarkdatensatzes (Visual Geometry Group, 2003) getestet (Abschnitt 6.1). Die Untersuchung der Leistungsfähigkeit des Gesamtsystems anhand eigener praxisnaher Bildsequenzen ist in Abschnitt 6.2 dargestellt. Die absolute Orientierung von Bildsequenzen auf der Grundlage von Kameras mit integriertem GPS wurde mittels eines Referenzsystems eingehend auf die erzielbare Genauigkeit untersucht (Abschnitt 6.3). Abschnitt 6.4 demonstriert die Orientierung von Bildern verschiedener Kameras anhand von Boden- und Luftaufnahmen.

6.1 Evaluation der invarianten / robusten Zuordnung

Anhand der Testbilddatensätze der Visual Geometry Group (2003) Oxford ist ein objektiver Vergleich zu bisherigen Verfahren zur Bildzuordnung möglich. Bei den Kriterien handelt es sich um Unschärfe (Abschnitt 6.1.1), größere Blickwinkeländerung (Abschnitt 6.1.2), Maßstabsunterschied in Verbindung mit Rotation um die Hauptachse (Abschnitt 6.1.3), Veränderung der Lichtverhältnisse (Abschnitt 6.1.4) und JPEG-Bildkomprimierung (Abschnitt 6.1.5). Es sind acht Testbildsätze mit jeweils sechs Bildern verfügbar, von denen jeweils eines als Master verwendet und paarweise den fünf übrigen Bildern des entsprechenden Satzes zugeordnet wird.

Mittels beigefügter Homographien ist es möglich, die Anzahl der korrekten Zuordnungen zu bestimmen. Als Entscheidungskriterium für eine korrekte Zuordnung wird die „Nächster-Nachbar“ Methode verwendet, wobei der Abstand zum korrespondierenden Punkt kleiner als fünf Pixel sein

muss. Bei der Bestimmung der Zuordnungen wurde wie in Kapitel 4 beschrieben die geometrische Plausibilität mittels robust geschätzter Epipolargeometrie mit einbezogen. Da für die Testbilder die intrinsischen Kameraparameter nicht zur Verfügung stehen, wurde der 7-Punkt Algorithmus (siehe Abschnitt 2.1), d.h. der unkalibrierte Fall, verwendet. Zur Beurteilung der Ansätze wurde jeweils die absolute Anzahl an korrekten Zuordnungen sowie der Anteil der korrekten Zuordnungen bestimmt. Des Weiteren wurde die Genauigkeit der Lokalisierung detailliert untersucht. Zu Vergleichszwecken wurden neben den selbst durchgeführten Tests für FASIAM, SIFT, SIFT & AKQZ und ASIFT die Ergebnisse für mehrere Verfahren aus (Mikolajczyk *et al.*, 2005) beigefügt. Die in den im Folgenden aufgeführten Diagrammen angegebenen Verfahren können wie folgt charakterisiert werden:

- Fast Accurate Scale Invariant Affine Matching (FASIAM): Eigener, neu entwickelter Ansatz (siehe Kapitel 4).

- Förstner & Affine Kleinste-Quadrate-Zuordnung (AKQZ): Dieser Ansatz verwendet Förstner Punkte (siehe Abschnitt 3.1), normierte Kreuzkorrelation und affine Kleinste-Quadrate-Zuordnung zur Bildzuordnung und es wird die geometrische Plausibilität einbezogen. Mit diesem Ansatz wurde zu Beginn gearbeitet, der Vergleich mit FASIAM verdeutlicht somit die Erweiterung der Leistungsfähigkeit.

- SIFT & AKQZ: Bei diesem Ansatz werden SIFT-Punkte und die Zuordnung über den SIFT-Deskriptor verwendet. Die zugeordneten Punkte werden durch eine affine Kleinste-Quadrate-Zuordnung verbessert, bei der, wie bei FASIAM, die verschiedenen Maßstäbe der Bildausschnitte berücksichtigt werden. Des Weiteren wird die geometrische Plausibilität einbezogen. Für die Extraktion der SIFT-Punkte wurde ein modifizierter Parametersatz verwendet, der zu etwas mehr Punkten führt, als beim Verfahren von Lowe (2004).

- Affine SIFT – ASIFT (siehe Abschnitt 3.3): Für den Test mit dem derzeit wohl besten Ansatz zur Bildzuordnung bei Blickpunktänderungen von Morel & Yu (2009) wurde eine von den Autoren zur Verfügung gestellte Demonstrationsversion verwendet.

- SIFT: Zum Vergleich der Ergebnisse von FASIAM mit dem häufig verwendeten SIFT-Operator wurde eine vom Autor zur Verfügung gestellte Demonstrationssoftware (Lowe, 2005) benutzt.

- Harris / Hessian Affine: Diese Ansätze basieren auf dem Harris Operator (Harris & Stephens, 1988). Sie wurden für eine robuste Zuordnung gegen Blickwinkeländerungen und Maßstabsunterschiede konzipiert.

- Maximally Stable Extremal Regions (MSER): Dies ist ein „Blob“ Operator, welcher verhältnismäßig große Bildregionen extrahiert, die durch lokale Binarisierung bestimmt werden (Matas *et al.*, 2002). Für die so erhaltenen Bildregionen erfolgt eine Normalisierung aller sechs Parameter der affinen Transformation.

- Viewpoint-Invariant Patches (VIP) von Wu *et al.* (2008). Dieser Ansatz setzt die Kenntnis der relativen Orientierung voraus und ermöglicht es auf diese Weise, auch über große Blickwinkeländerungen korrekte Punktkorrespondenzen zu bestimmen. Es liegen lediglich für einen Bilddatensatz (Wand) Ergebnisse vor.

Für die drei übrigen im Weiteren verwendeten Ansätze Intensity Based Regions (IBR), Edge Based Regions (EBR), Affine invariant salient region detector (Salient) gilt im Wesentlichen dasselbe wie für MSER: Sie wurden eher für Anwendungen im Bereich der Objektextraktion konzipiert und sind für die Fragestellungen dieser Arbeit nur insofern relevant, weil sie von vielen anderen Arbeiten zu Vergleichszwecken aufgeführt werden.

6.1.1 Unschärfe

Die Auswirkungen von Unschärfe auf die verschiedenen Verfahren zur Bildzuordnung werden an den Bildsätzen Motorräder (siehe Abb. 6.1 und 6.2) und Bäume (siehe Abb. 6.4 und 6.5) aufgezeigt. Die zugehörigen Bilder wurden hierfür durch Veränderung der Fokussierung zunehmend unschärfer. Kameraposition und Blickrichtung bleiben nahezu unverändert. Die Bildsätze unterscheiden sich dahingehend, dass der Motorrad-Datensatz eine Szene mit vielen verhältnismäßig großen, eintönigen Regionen zeigt, während die Szene des Baum-Datensatzes überwiegend kleine, feine Strukturen aufweist.

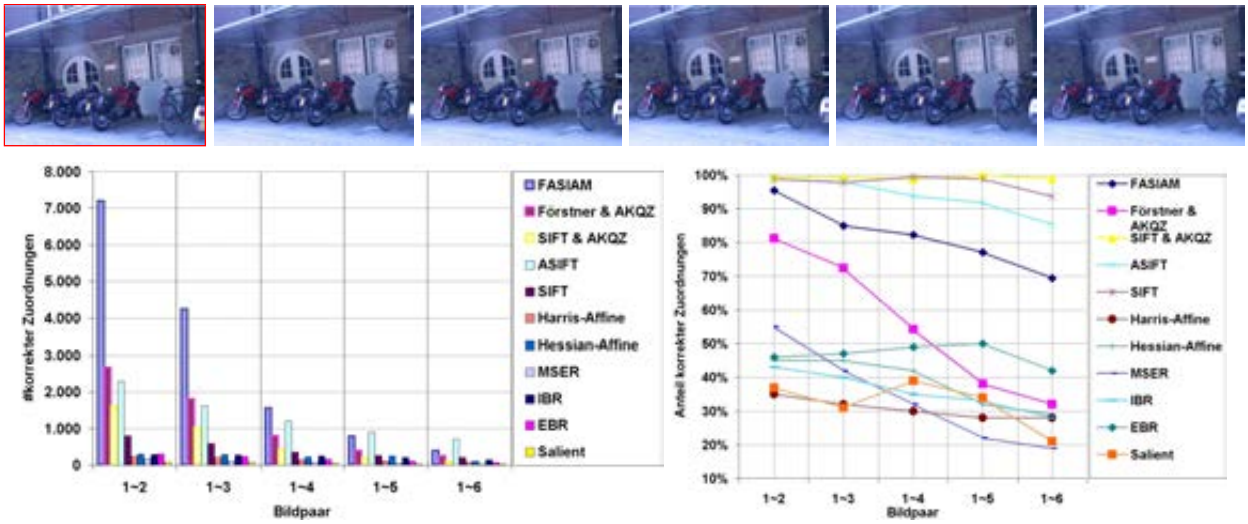


Abbildung 6.1: Unten: Absolute Anzahl (# – unten links), sowie Anteil (% – unten rechts) an korrekten Zuordnungen für den Bildsatz Motorräder – Unschärfe (oben).

Die absolute Anzahl der korrekten Zuordnungen fällt für FASIAM bei größerer Unschärfe stark ab. Das ist dadurch zu erklären, dass FASIAM die eingehenden Bilder möglichst wenig glättet. Dies ist bei Maßstabsunterschieden vorteilhaft, für diesen Störeinfluss aber nachteilhaft. Obgleich der Abfall des Anteils an korrekten Zuordnungen von über 95% auf 70% vergleichsweise groß ist,

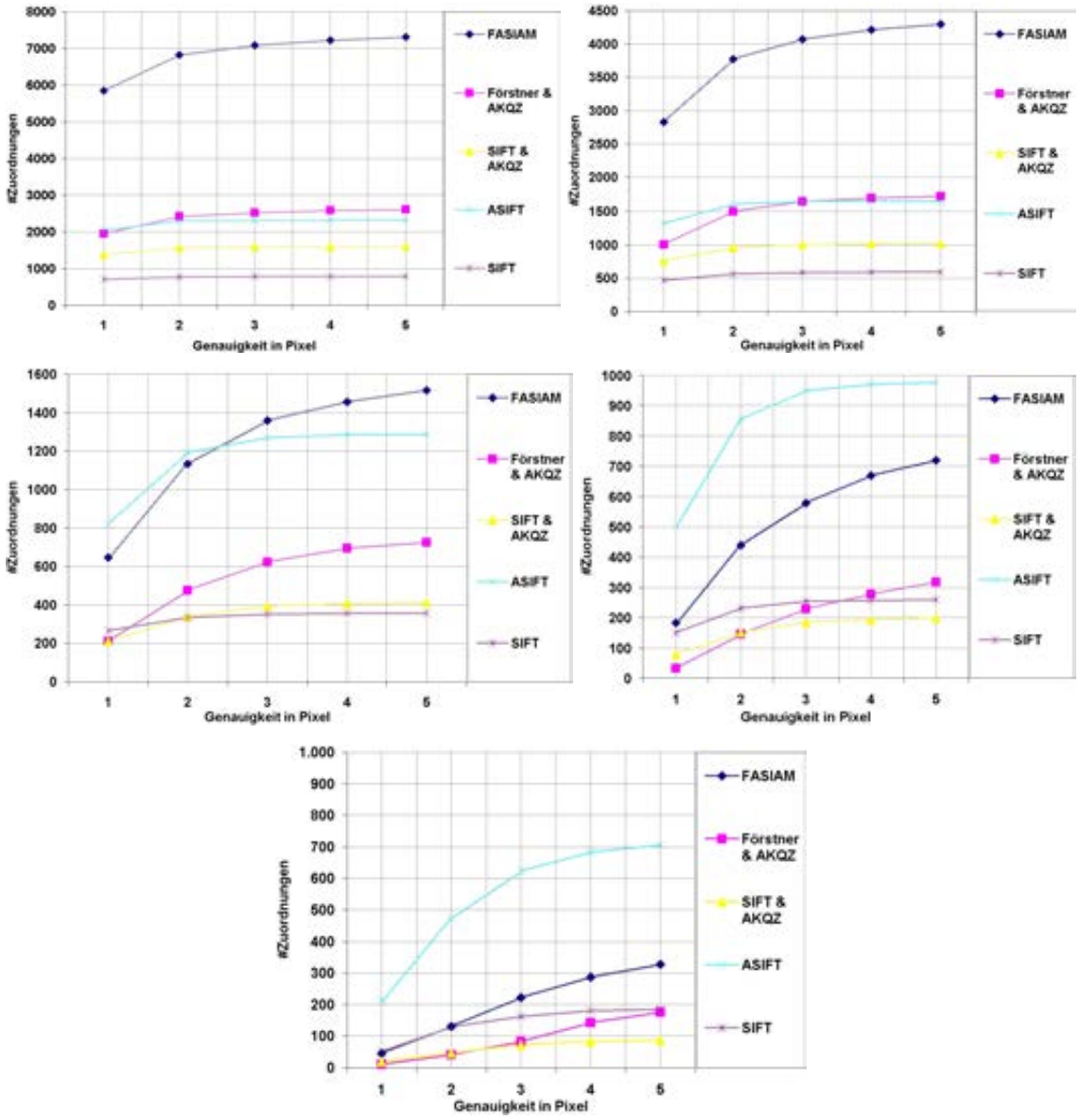


Abbildung 6.2: Genauigkeiten für den Bildsatz Motorräder – Unschärfe: Von oben links nach unten: Bildpaar $1 \sim 2$ bis $1 \sim 6$.

ist eine robuste Parameterschätzung unter Verwendung von FASIAM Zuordnungen auch für das schwierigste Bildpaar (siehe Abb. 6.3) problemlos möglich. Für die ersten beiden Bildpaare erzielt FASIAM sehr genaue Ergebnisse. Dies ist an der großen Zahl der Zuordnungen mit einer Ge-

nauigkeit von bis zu einem Pixel erkennbar. Sowohl die Zahl der genauen Zuordnungen als auch die Anzahl insgesamt fallen ab dem Bildpaar $1 \sim 4$ stark ab. Einzig ASIFT liefert für die letzten beiden Paare noch gute Ergebnisse. Eine Optimierung bezüglich Zuordnung bei Unschärfe sollte weiterverfolgt werden, da auch unscharfe Bilder in einer Bildmenge bei einer praktischen Nutzung als Normalfall angesehen werden müssen.

Abbildung 6.3: Bildsatz Motorräder-Unschärfe: 406 mittels FASIAM generierte korrekte Zuordnungen für das Bildpaar $1 \sim 6$.

Der im Folgenden untersuchte Bildsatz Bäume ist ebenfalls für den Test auf Robustheit gegen Unschärfe vorgesehen. Die sich wiederholenden und feinen Strukturen erschweren die Bildzuordnung bei Unschärfe aber erheblich.

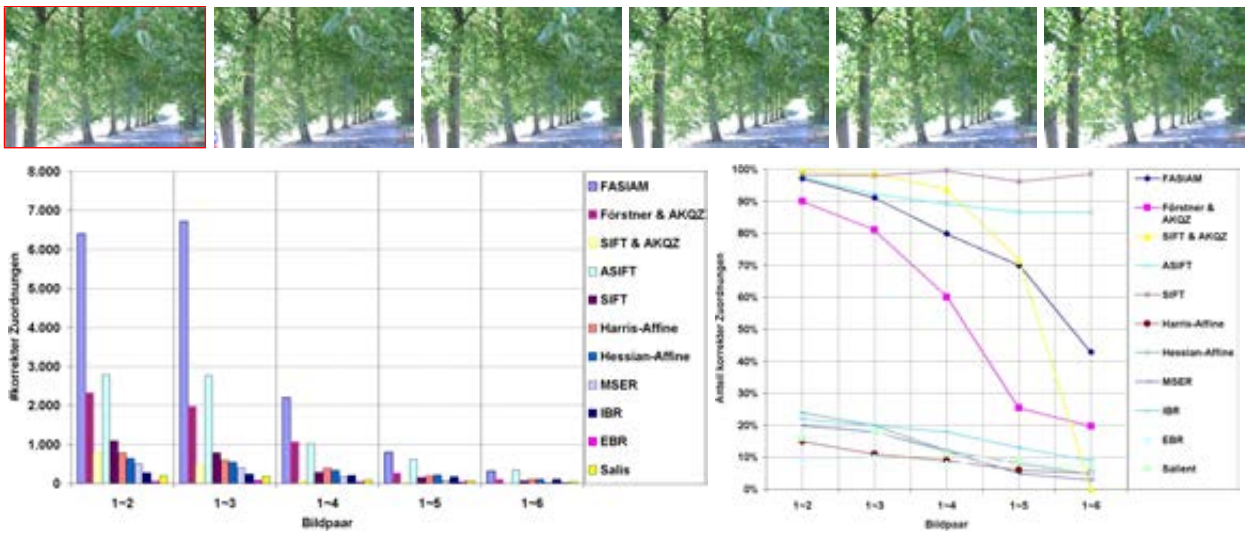


Abbildung 6.4: Absolute Anzahl (# – unten links) und Anteil (% – unten rechts) der korrekten Zuordnungen für den Bildsatz Bäume – Unschärfe (oben).

Bei der Anzahl und dem Anteil der Zuordnungen sind die Ergebnisse aller Verfahren, außer dem eigenen Ansatz, ab dem dritten Bildpaar deutlich schlechter, als beim Bildsatz Motorräder. FASIAM liefert bis einschließlich zum Bildpaar $1 \sim 5$ gute Ergebnisse. Für das sechste Bildpaar fällt der Anteil der korrekten Zuordnungen auf etwas über 40% verhältnismäßig stark ab.

Bei den Ergebnissen für die Genauigkeiten ist zu erkennen, dass verglichen mit dem Bildsatz Motorräder, die Zuordnungen deutlich früher, nämlich ab dem zweiten Bildpaar ungenauer werden, was durch die abgebildete Vegetation erklärbar ist. FASIAM erzielt bis zum fünften Bildpaar gute Ergebnisse, nur für das sechste Bildpaar weist ASIFT bessere auf. Aufgrund der wenigen genauen Zuordnungen beim letzten Bildpaar, war eine genaue Parameterschätzung mittels FASIAM nicht möglich. Es ist offensichtlich, dass ein geringes Maß an Unschärfe auch bei Bildern mit feinen Strukturen bei allen getesteten Verfahren kaum Auswirkungen auf die Möglichkeiten zur Bildzuordnung hat. Der Anteil an genauen Zuordnungen nimmt jedoch mit Zunahme der Unschärfe bei allen Verfahren schnell ab.

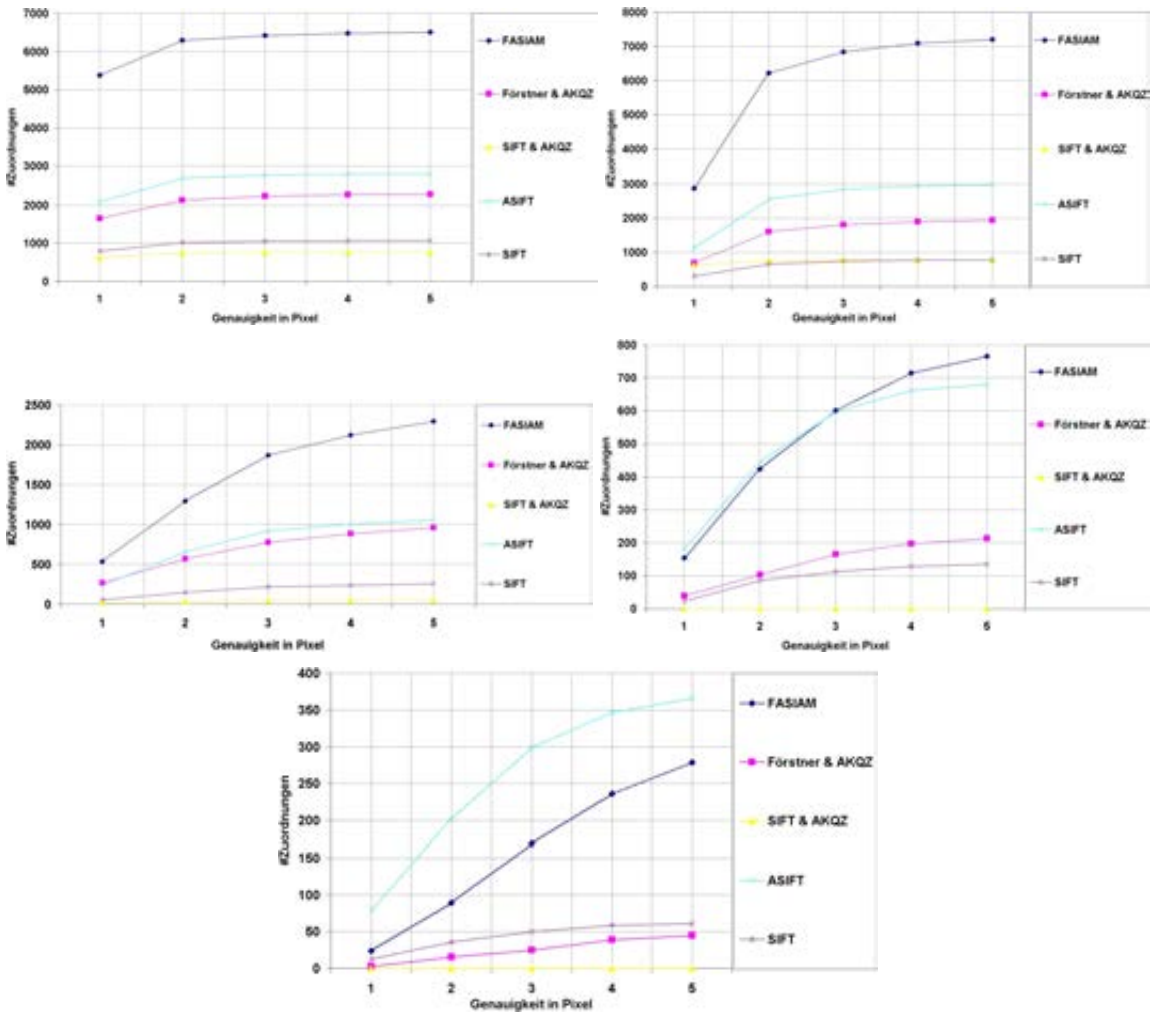


Abbildung 6.5: Genauigkeiten für den Bildsatz Bäume – Unschärfe: Von oben links nach unten: Bildpaar $1 \sim 2$ bis $1 \sim 6$

6.1.2 Blickwinkeländerung

Die Auswirkungen von Blickwinkeländerungen auf die verschiedenen Verfahren zur Bildzuordnung werden an den Bildsätzen Wand (siehe Abb. 6.6 und Abb. 6.7) und Grafity (siehe Abb. 6.9 und Abb. 6.10) getestet. Beim Datensatz Wand wurde bei den zugehörigen Bildern zunächst nahezu senkrecht auf eine Ziegelsteinwand geblickt und dann für jedes weitere Bild der Blickwinkel vergrößert. Die Bilder weisen vorwiegend feine und zyklisch wiederkehrende Strukturen auf. Die

Bilder des Datensatzes Grafity zeigen eine mit Lackspray bemalte Wand, so dass sie vorwiegend große eintönige Regionen aufweisen. Zudem wurde bei den Aufnahmen zusätzlich zur Blickwinkeländerung die Kamera gedreht.

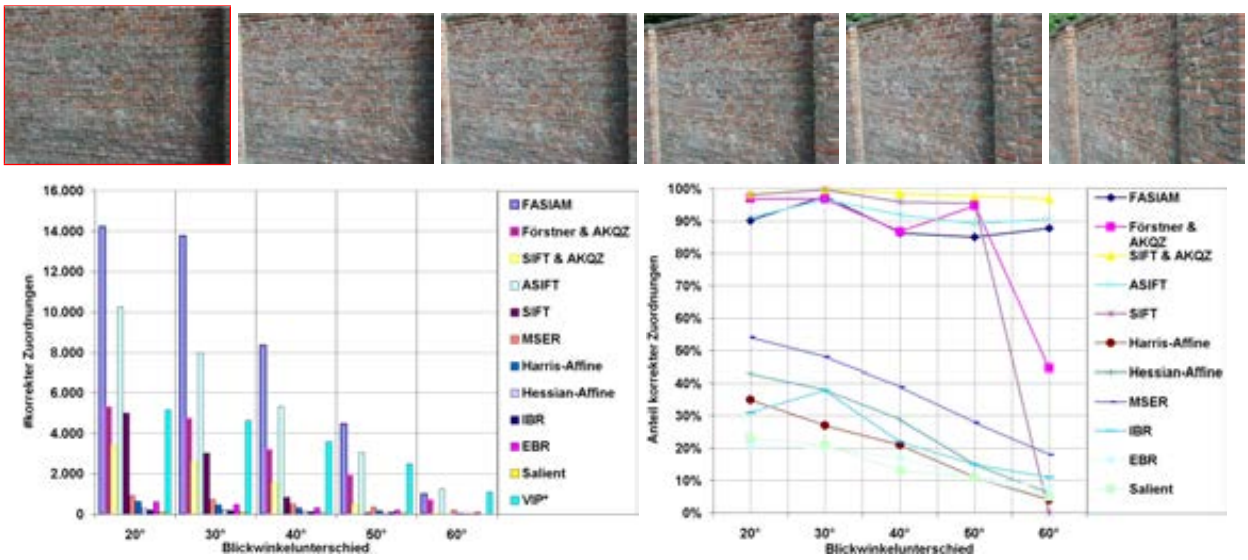


Abbildung 6.6: Absolute Anzahl (# – unten links) und Anteil (% – unten rechts) an korrekten Zuordnungen für den Bildsatz Wand – Blickwinkeländerung (oben). Für diesen Bildsatz sind zusätzlich zu den erläuterten Verfahren Testergebnisse für die Viewpoint-Invariant Patches – VIP (Wu *et al.*, 2008) verfügbar, welche die Kenntnis der relativen Orientierung voraussetzen.

Abgesehen vom größten Blickwinkelunterschied (siehe Abb. 6.8) erzielt FASIAM für den Bildsatz Wand die besten Ergebnisse. Es werden selbst VIP (Wu *et al.*, 2008) und ASIFT übertroffen. Vermutlich ist dies auf die stark texturierte Struktur der Ziegelsteinwand zurückzuführen, welche für die Bildzuordnung basierend auf normierter Kreuzkorrelation sehr vorteilhaft ist. Dies erklärt auch die guten Ergebnisse für die auf Förstner-Punkten basierende Methode. Bezüglich des Anteils der korrekten Zuordnungen ist deutlich zu erkennen, dass bei allen Bildpaaren für FASIAM ein hinreichend hoher Anteil an korrekten Zuordnungen erzielt wurde. Anzumerken ist, dass dies im Wesentlichen durch die Einbeziehung der geometrischen Plausibilität und der für das Verfahren vorteilhaften Textur möglich ist. Ähnlich beschaffene Szenen sind aber in der Praxis häufig zu erwarten.

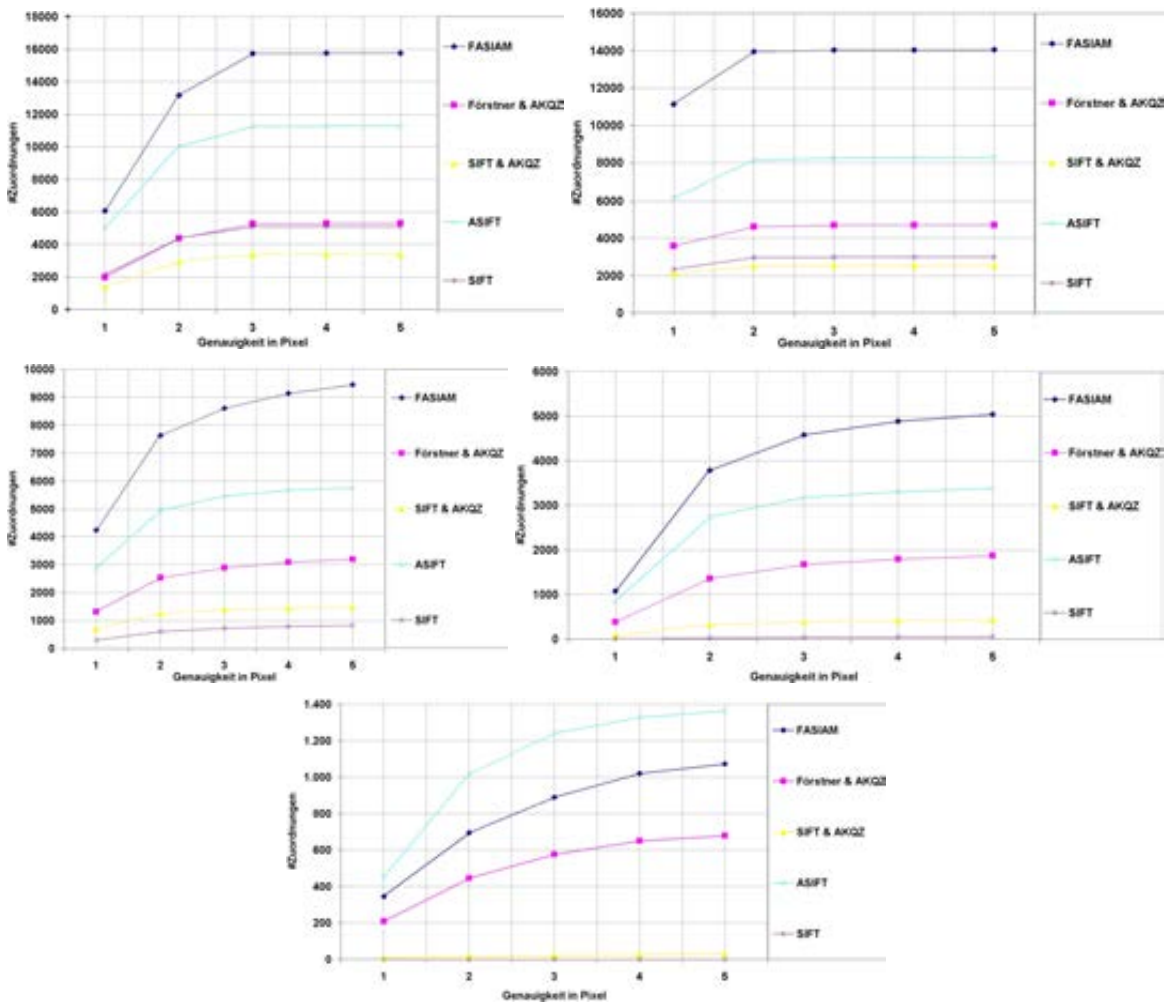


Abbildung 6.7: Genauigkeiten für den Bildsatz Wand – Blickwinkeländerung: Von oben links nach unten: Bildpaar $1 \sim 2$ bis $1 \sim 6$.

FASIAM und ASIFT weisen deutlich bessere Ergebnisse auf, als die übrigen Verfahren. Auffallend ist der verhältnismäßig hohe Anteil an etwas ungenaueren Zuordnungen zwischen ein und zwei Pixel.

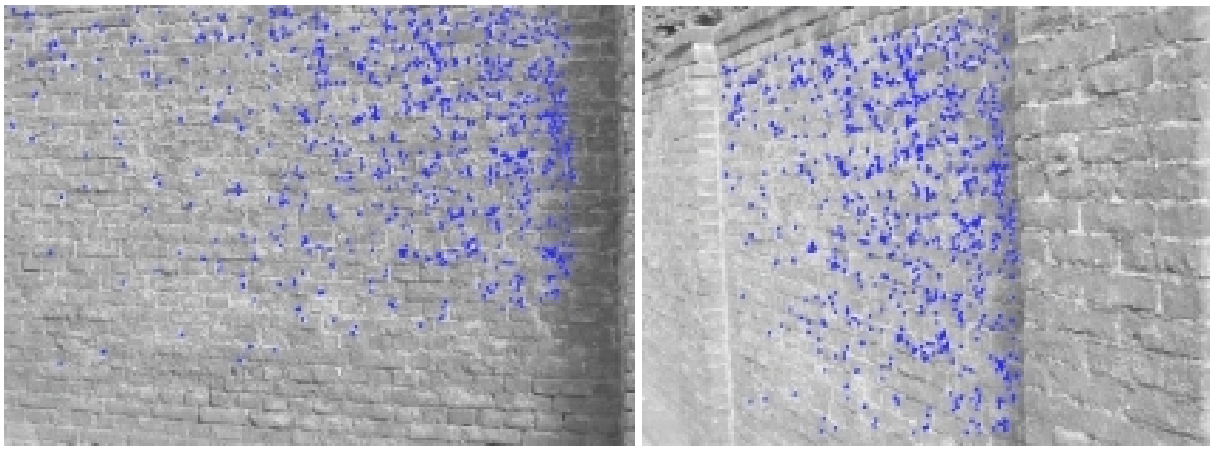


Abbildung 6.8: Bildsatz Wand – Blickwinkeländerung: 894 mittels FASIAM generierte korrekte Zuordnungen für das Bildpaar $1 \sim 6$ (Blickwinkelunterschied 60°).

An den Ergebnissen für den Bildsatz Grafity ist deutlich erkennbar, dass fast alle Verfahren bei diesem an ihre Grenzen stoßen.

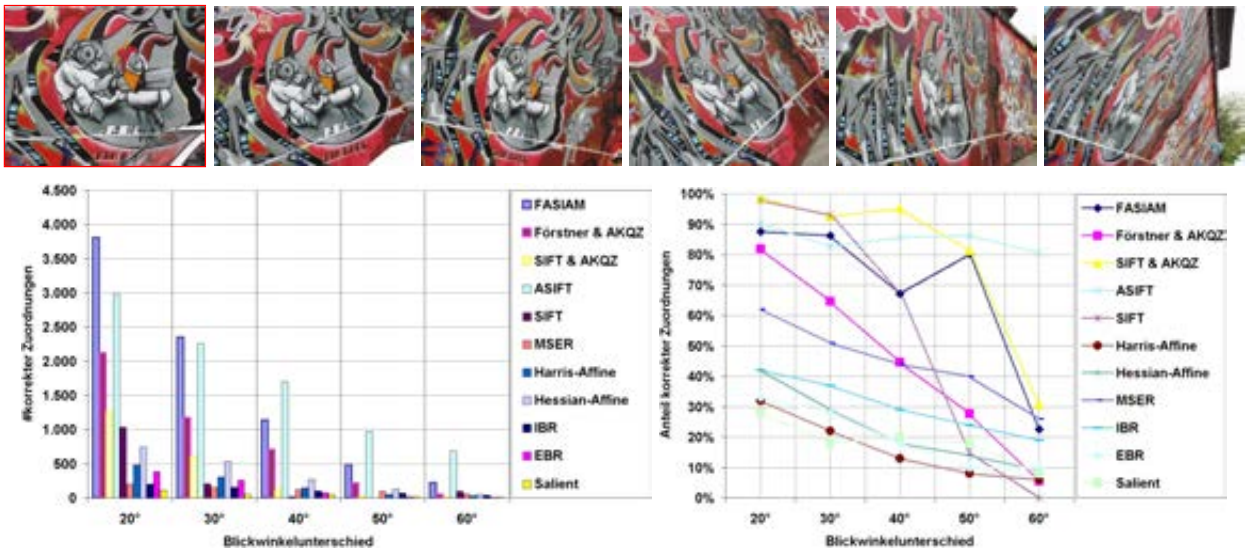


Abbildung 6.9: Absolute Anzahl (# – unten links) und Anteil (% – unten rechts) an korrekten Zuordnungen für den Bildsatz Grafity – Blickwinkeländerung (oben).

Insbesondere bei den starken Blickwinkeländerungen über 50° und 60° (Bildpaar $1 \sim 5$ und $1 \sim 6$) liefert nur noch ASIFT gute Resultate. Die Ursache liegt in der Beschaffenheit der Szene, welche verhältnismäßig große eintönige Bildregionen aufweist, die per se zu einem hohen Anteil an Fehlzuordnungen führen, was durch die perspektive Verzerrung, die Rotation um die Hauptachse und die kleiner werdende Überlappung nochmals deutlich verschärft wird. Beim Anteil der korrekten Zuordnungen wird der Unterschied zum Bildsatz Wand deutlich.

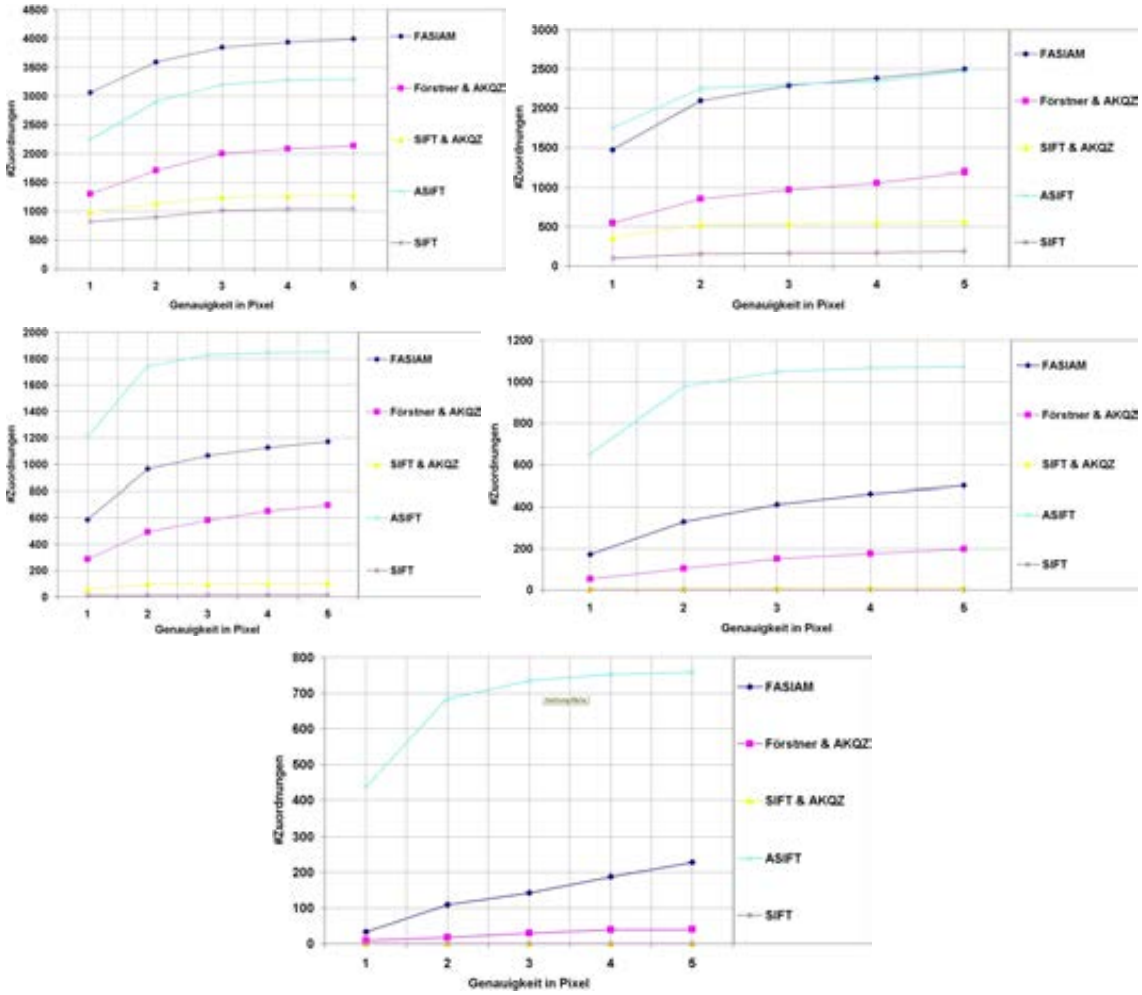


Abbildung 6.10: Genauigkeiten für den Bildsatz Grafity – Blickwinkeländerung: Von oben links nach unten: Bildpaar $1 \sim 2$ bis $1 \sim 6$.

Die Werte für FASIAM fallen deutlich früher und stärker ab. Eine verlässliche robuste Parame-

terschätzung ist für FASIAM nur bis zu einem Blickwinkelunterschied von 50° möglich. Auch bei den Ergebnissen für die Genauigkeiten ist erkennbar, dass die Bildzuordnung für diesen Bildsatz sehr schwierig ist. Ab dem dritten Bildpaar ist der Anteil an ungenauen Zuordnungen für alle getesteten Verfahren hoch. Für das Bildpaar $1 \sim 5$ liefern nur noch FASIAM und ASIFT verwertbare Ergebnisse. Für das letzte Bildpaar mit einem Blickwinkelunterschied von 60° ist nur noch ASIFT geeignet.

6.1.3 Maßstabsunterschied und Rotation um die Hauptachse

Die Auswirkungen von Maßstabsunterschieden und Rotation um die Hauptachse auf die verschiedenen Verfahren zur Bildzuordnung werden an den Bildsätzen Rinde (siehe Abb. 6.11 und Abb. 6.12) und Boot (siehe Abb. 6.14 und Abb. 6.15) getestet. Bei den zugehörigen Bildern wurde zwischen den Aufnahmen immer weiter „weggezoomt“, so dass der Maßstabsunterschied zum ersten Bild größer wird. Zusätzlich wurde die Kamera um die Hauptachse gedreht. Beim Bildsatz Boot handelt es sich um Grauwertbilder, was die Bildzuordnung zusätzlich erschwert. Die begrenzte Robustheit der SIFT-Zuordnung gegen Maßstabsunterschiede wird an diesem Bildsatz besonders offensichtlich.

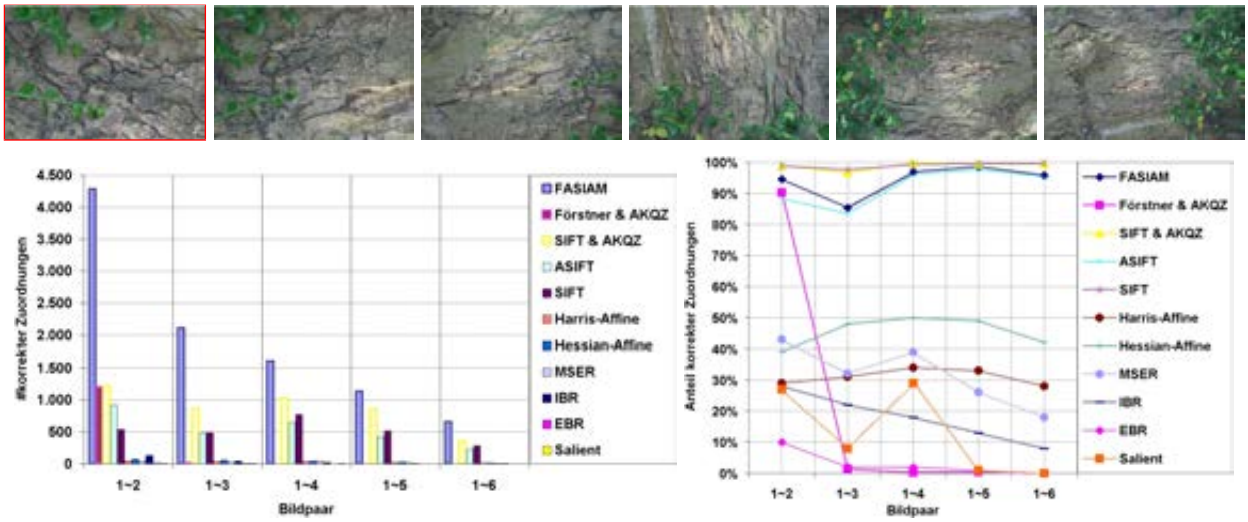


Abbildung 6.11: Absolute Anzahl (# – unten links) und Anteil (% – unten rechts) an korrekten Zuordnungen für den Bildsatz Rinde (oben) – Maßstabsunterschied und Rotation um die Hauptachse.

Abb. 6.11 zeigt besonders deutlich die Schwäche des auf Förstner-Punkten basierenden Verfahrens sowie die deutliche Verbesserung durch FASIAM, bei dem auch noch für das Bildpaar $1 \sim 6$ viele korrekte Zuordnungen erzeugt werden (siehe Abb. 6.13). Auch für den Anteil der korrekten Zuordnungen ist ein signifikanter Unterschied zwischen der auf Förstner-Punkten basierenden Methodik und FASIAM erkennbar.

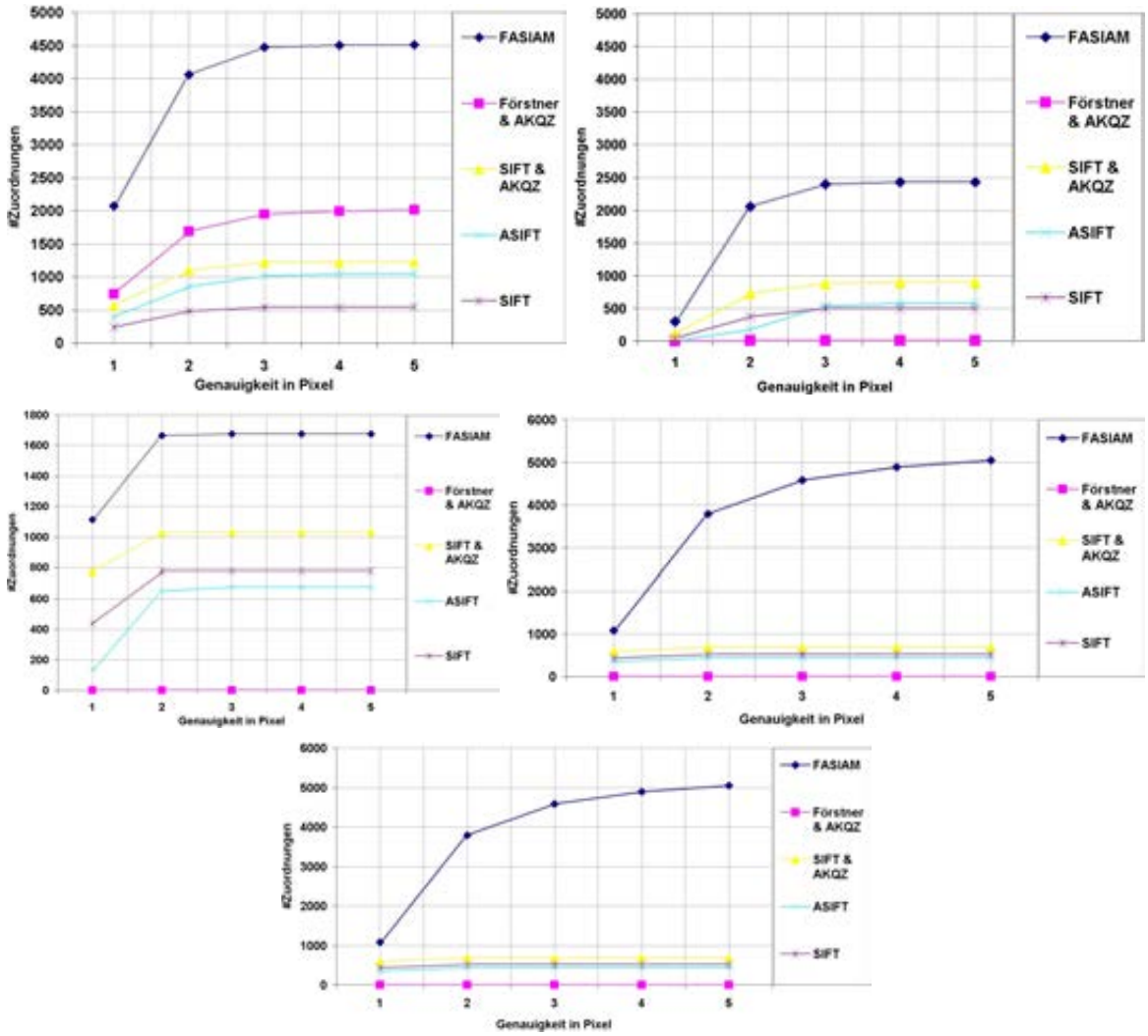


Abbildung 6.12: Genauigkeiten für den Bildsatz Rinde – Maßstabsunterschied und Rotation um die Hauptachse: Von oben links nach unten: Bildpaar $1 \sim 2$ bis $1 \sim 6$.

Des Weiteren wird deutlich, warum die zum Vergleich herangezogenen Verfahren Harris / Hes-

sian Affine, MSER, etc. kaum für eine praxisnahe Anwendung geeignet erscheinen: Der Anteil der korrekten Zuordnungen ist sehr niedrig. Zusätzlich ist anzumerken, dass fast alle Verfahren beim Bildpaar $1 \sim 3$ einen schlechteren Wert erzielen, als beim nächst schwierigeren Paar $1 \sim 4$.

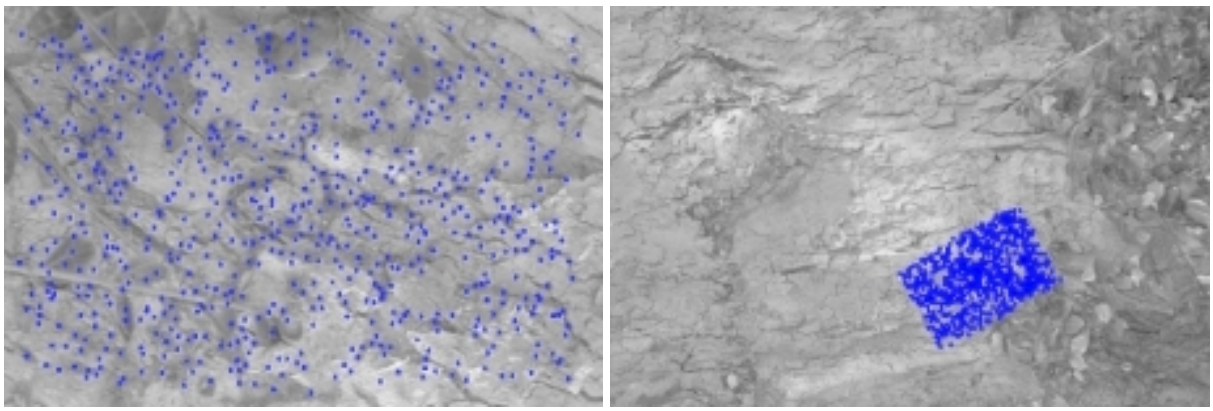


Abbildung 6.13: Bildsatz Rinde: 664 mittels FASIAM generierte korrekte Zuordnungen für das Bildpaar $1 \sim 6$.

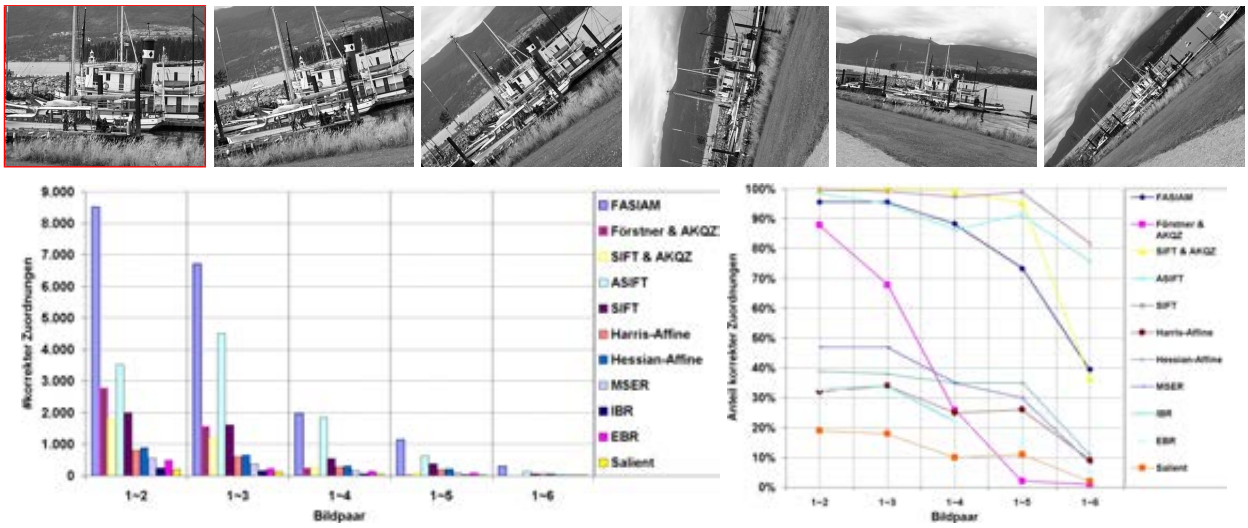


Abbildung 6.14: Absolute Anzahl (# – unten links) und Anteil (% – unten rechts) der korrekten Zuordnungen für den Bildsatz Boot (oben) – Maßstabsunterschied und Rotation um die Hauptachse.

Ursache hierfür scheint zu sein, dass die dem Bildpaar $1 \sim 3$ zugehörige Homographie ungenau ist. Bei den von FASIAM erzielten Ergebnissen fällt eine hohe Anzahl an Zuordnungen auf, bei denen der Anteil der eher ungenaueren aber verhältnismäßig hoch ist. Auch hier weisen die Ergebnisse auf eine ungenaue gegebene Homographie für das Bildpaar $1 \sim 3$ hin. Der im Folgenden untersuchte Bildsatz Boot weist als einziger Grauwertbilder auf. Für diesen Bildsatz nimmt die Anzahl an korrekten Zuordnungen für alle getesteten Verfahren schnell ab.

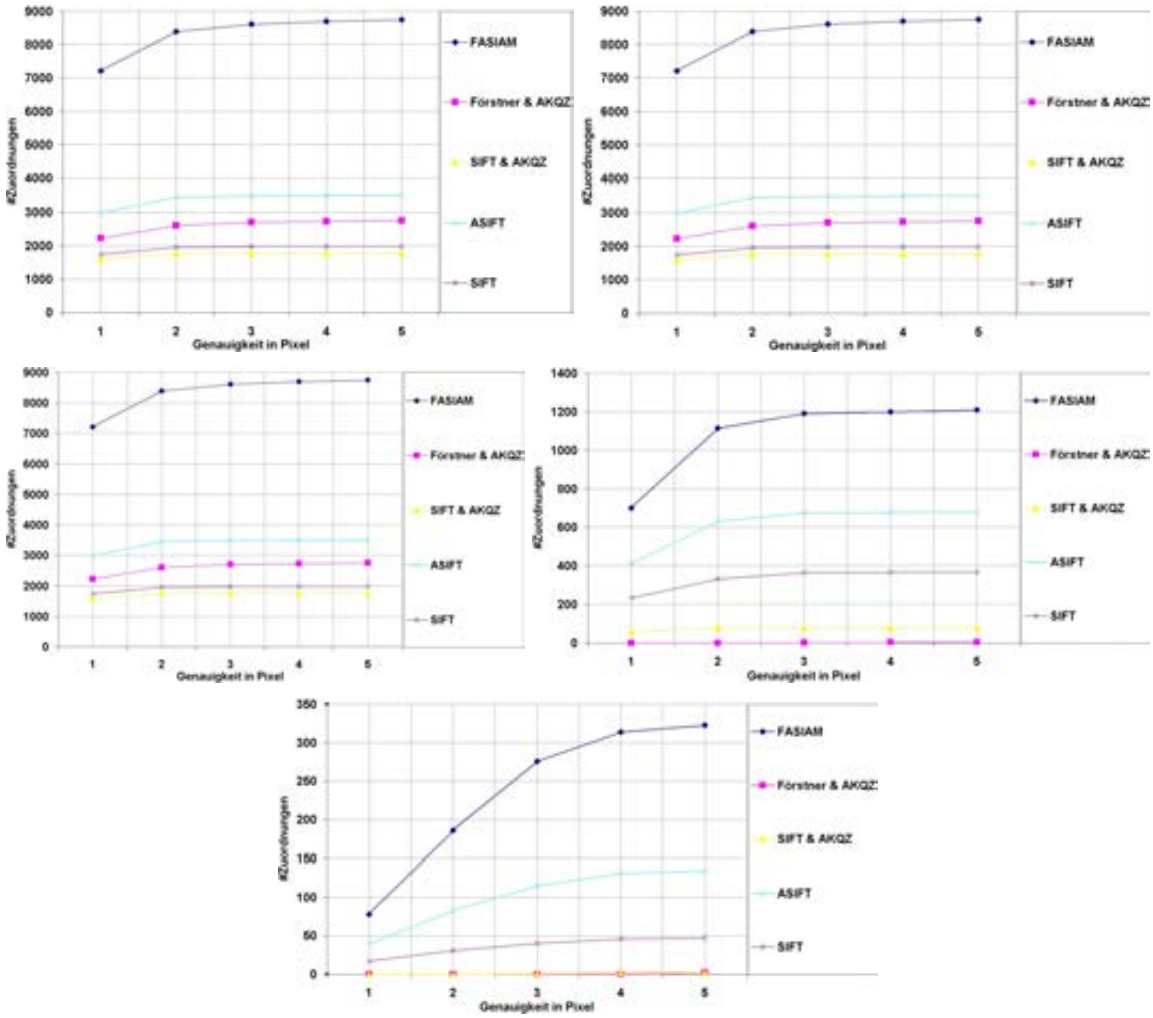


Abbildung 6.15: Genauigkeiten für den Bildsatz Boot – Maßstabsunterschied und Rotation um die Hauptachse: Von oben links nach unten: Bildpaar $1 \sim 2$ bis $1 \sim 6$.

Für das letzte Bildpaar werden kaum noch korrekte Zuordnungen gefunden. Die höchste Anzahl weist auch hier FASIAM auf. Auch für den Anteil der korrekten Zuordnungen ist ein signifikanter Unterschied zwischen der ursprünglichen, auf Förstner-Punkten basierenden Methode und FASIAM erkennbar. Beim Bildpaar $1 \sim 6$ ist der Anteil an korrekten Zuordnungen auch für FASIAM eher niedrig. FASIAM erzielt bei diesem Bildsatz auch bei den Genauigkeiten durchgehend gute bis sehr gute Ergebnisse. Lediglich beim Bildpaar $1 \sim 6$ steigt der Anteil an ungenauen Zuordnungen deutlich an.

6.1.4 Änderung der Lichtverhältnisse

Die Auswirkungen der Änderung der Lichtverhältnisse auf die verschiedenen Verfahren zur Bildzuordnung werden an einem Bildsatz getestet, bei dem zwischen den Aufnahmen die Belichtungseinstellung der Kamera geändert wurden, so dass die Szene stetig dunkler erscheint (siehe Abb. 6.16 und Abb. 6.17).

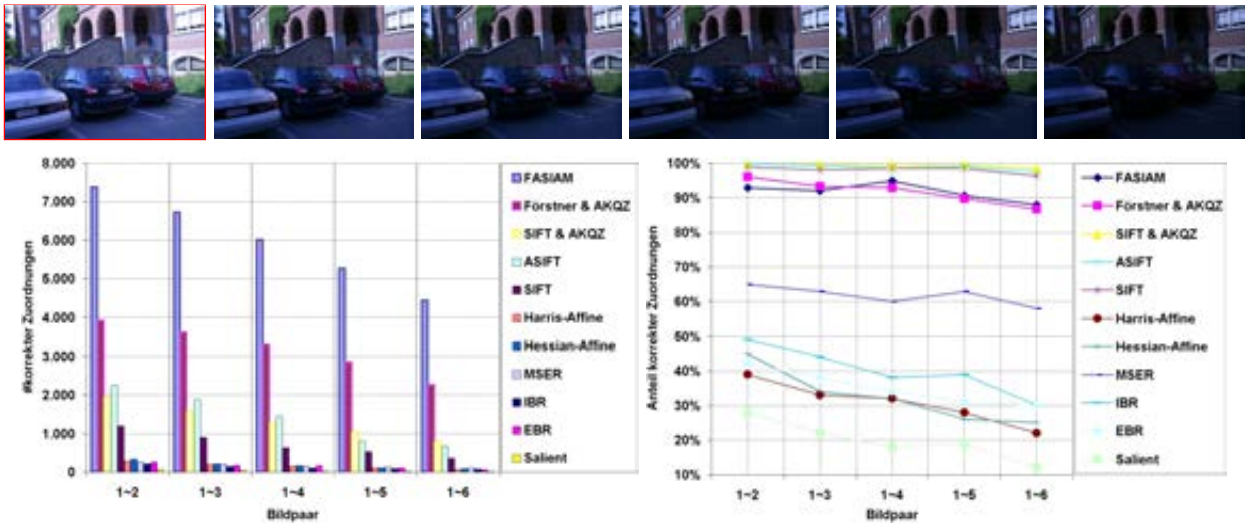


Abbildung 6.16: Absolute Anzahl (# – unten links) und Anteil (% – unten rechts) an korrekten Zuordnungen für den Bildsatz Änderung Lichtverhältnisse (oben).

Abb. 6.16 verdeutlicht die Invarianz der normierten Kreuzkorrelation gegen gleichmäßige Veränderung der Lichtverhältnisse. Dies ist an den Ergebnissen für FASIAM und der auf Förstner-Punkten basierenden Methodik erkennbar. Beide generieren über alle Bildpaare deutlich bessere

Ergebnisse als alle anderen Verfahren. Zusätzlich zur sehr hohen Anzahl an Zuordnungen liegt auch der Anteil der korrekten Zuordnungen fast durchgehend über 90%. Das neue Verfahren ist somit sehr robust gegen diesen Störeinfluss, welcher für eine praktische Anwendung als Normalfall angesehen werden muss.

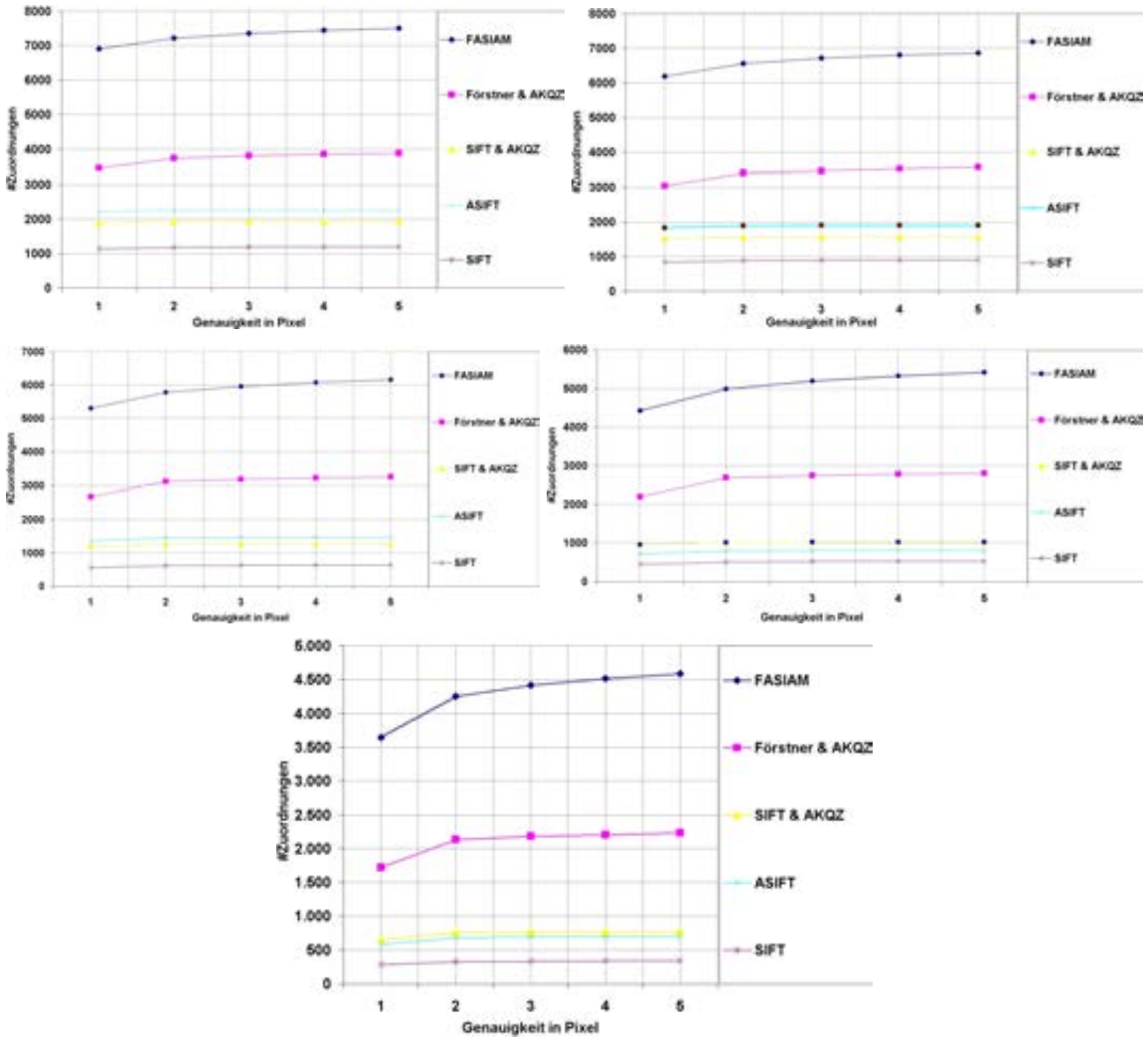


Abbildung 6.17: Genauigkeiten für den Bildsatz Änderung Lichtverhältnisse.

Die auf der Grundlage von normierter Kreuzkorrelation generierten Ergebnisse sind auch hinsichtlich ihrer Genauigkeit sehr gut. Lediglich beim Bildpaar $1 \sim 6$ (siehe Abb. 6.18) tritt ein nennenswerter Anteil an ungenaueren Zuordnungen auf.

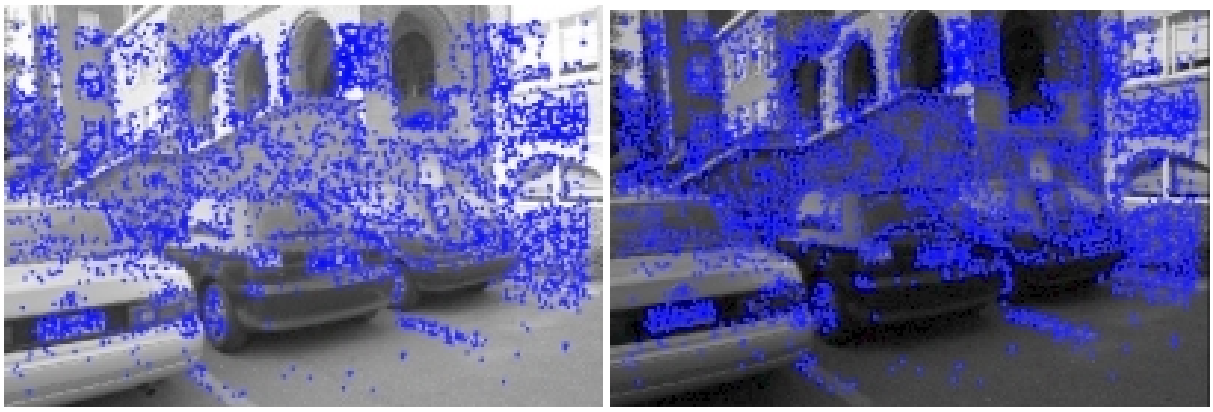


Abbildung 6.18: Bildsatz Änderung Lichtverhältnisse: 4.379 mittels FASIAM generierte korrekte Zuordnungen für das Bildpaar $1 \sim 6$.

6.1.5 JPEG-Bildkomprimierung

Die Auswirkungen der JPEG-Bildkomprimierung auf die verschiedenen Verfahren zur Bildzuordnung werden an einem Bildsatz getestet, bei dem der Komprimierungsgrad stetig erhöht wurde.

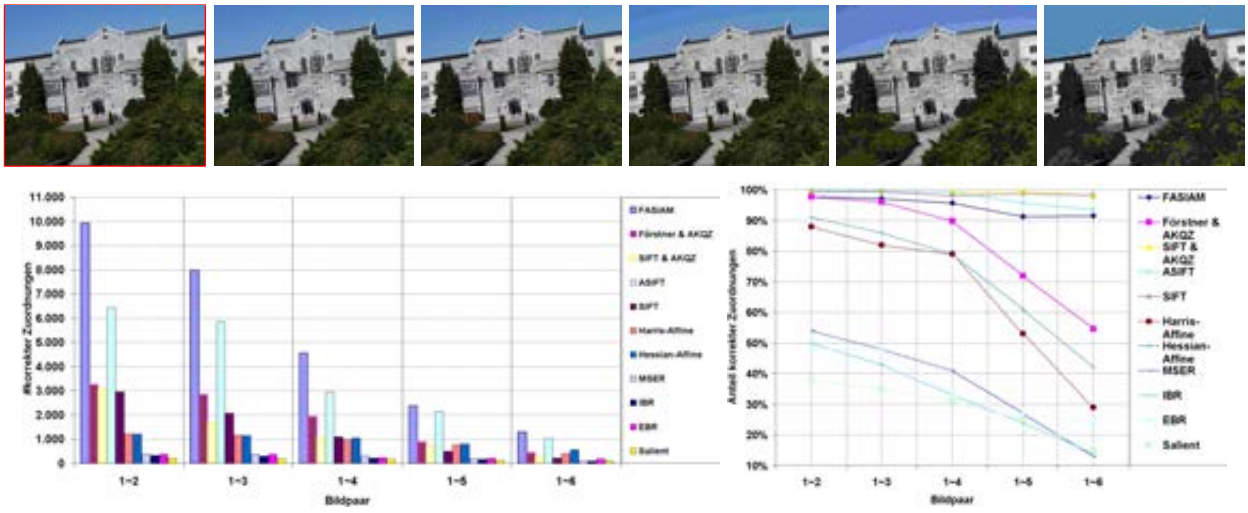


Abbildung 6.19: Absolute Anzahl (# – unten links) und Anteil (% – unten rechts) an korrekten Zuordnungen für den Bildsatz JPEG-Bildkomprimierung (oben).

Kameraposition und Blickrichtung bleiben unverändert. Sowohl für die Anzahl als auch beim Anteil der korrekten Zuordnungen generiert FASIAM sehr gute Ergebnisse. Abgesehen von den Verfahren, die auf SIFT-Zuordnung basieren, generiert nur FASIAM durchgehend einen sehr hohen Anteil an korrekten Zuordnungen.

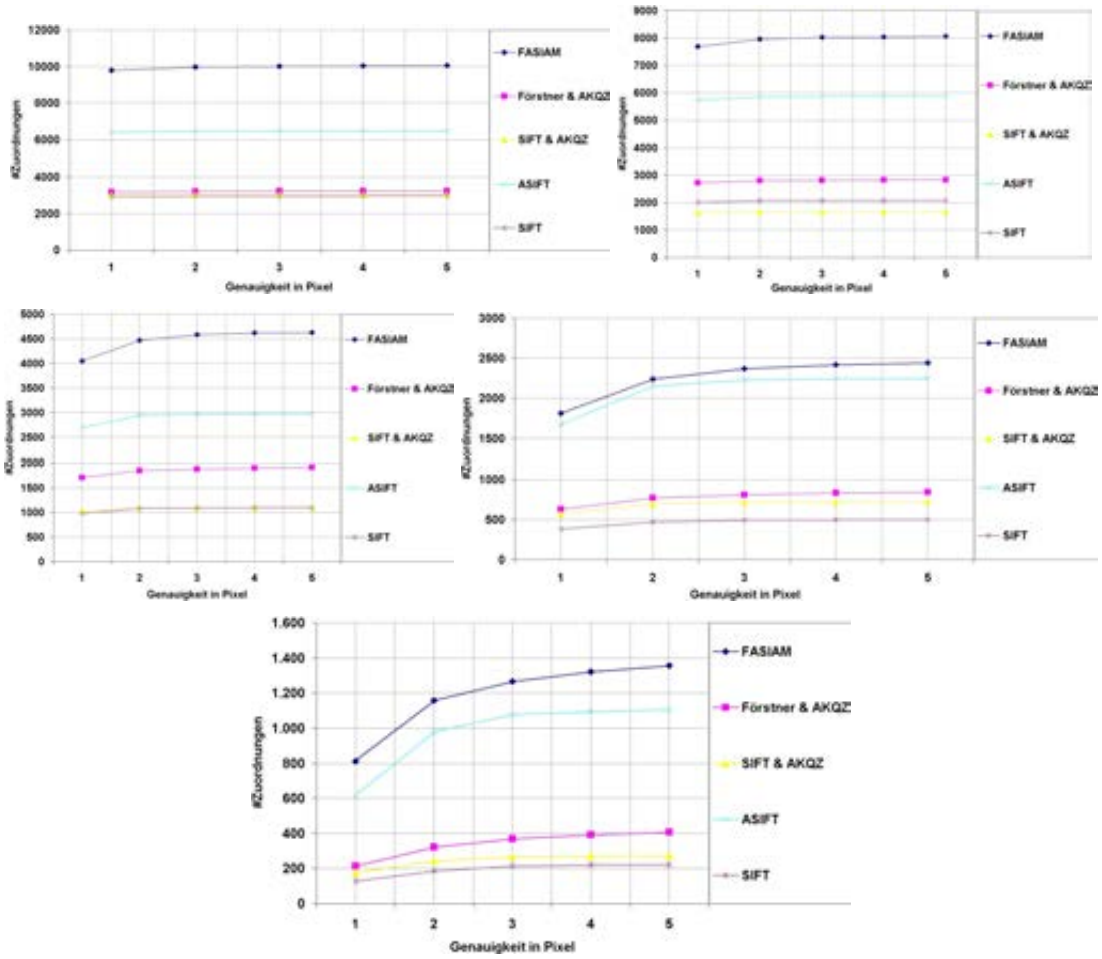


Abbildung 6.20: Genauigkeiten für den Bildsatz JPEG-Bildkomprimierung: Von oben links nach unten: Bildpaar $1 \sim 2$ bis $1 \sim 6$.

Obgleich die Zahl der Zuordnungen sehr groß ist, sind die Genauigkeiten für alle Verfahren sehr hoch. Lediglich beim Bildpaar $1 \sim 6$ tritt ein größerer Anteil an ungenauen Zuordnungen auf. Zu beachten ist, dass im Gegensatz zu allen anderen Bildsequenzen die Zuordnungen fehlerfrei verifiziert werden können, da die Perspektive nicht verändert wird.

6.1.6 Beurteilung der Ansätze zur Bildzuordnung

Die von Mikolajczyk *et al.* (2005) untersuchten Ansätze Harris-Affine, Hessian-Affine, MSER, IBR, EBR und Salient weisen nahezu durchgehend eine geringe Anzahl an Zuordnungen und einen Anteil an korrekten Zuordnungen auf, der unter 50% liegt. Lediglich für die JPEG-Komprimierung liegen bessere Ergebnisse vor. Für eine praktische Nutzung von Bildern mit größeren Blickwinkel-, Maßstabs- und Intensitätsänderung erscheinen diese Verfahren daher nur begrenzt geeignet. SIFT und ASIFT weisen aufgrund der Deskriptor-basierten Zuordnung durchgehend sehr hohe Anteile an korrekten Zuordnungen auf. Die Zahl der Zuordnungen sinkt bei vielen Kriterien jedoch rapide, d.h. es ist ein harter Übergang zwischen guten und sehr schlechten Ergebnissen erkennbar. Für die SIFT-Zuordnung lässt sich somit im Wesentlichen sagen, dass die Bildzuordnung entweder sehr gut oder gar nicht möglich ist. Im Gegensatz dazu weisen die Bildausschnitt-basierten Verfahren unter Verwendung der normierten Kreuzkorrelation einen weichen Übergang zwischen sehr guten und schlechten Ergebnissen auf. Dies eröffnet gute Nutzungsmöglichkeiten für schwierige Aufnahmekonfigurationen. Auf der Grundlage von Förstnerpunkten und normierter Kreuzkorrelation können gute Ergebnisse erzielt werden, wenn nur geringe Maßstabsunterschiede auftreten. Die Ergebnisse für den in dieser Arbeit entwickelten Ansatz belegen, dass dieses Problem unter Verwendung von Maßstabsraumextremwerten gelöst werden kann. D.h., FASIAM liefert für alle Kriterien gute bis sehr gute Ergebnisse.

6.2 Experimente mit dem Gesamtsystem

Im Weiteren wird davon ausgegangen, dass eine ungefähre Kalibrierung aller verwendeten Kameras entweder aus vorangegangenen Experimenten oder aus der Exchangeable Image File Format (Exif)-Information der Bilder zusammen mit weiterer Information wie der Pixelgröße bekannt ist. Der einfachste und im vorliegenden Fall am häufigsten verwendete Anwendungsfall für die Orientierung von Bildmengen sind Sequenzen von Bodenaufnahmen. Man nimmt hierbei Bilder auf, während man sich meist ungefähr senkrecht zur Aufnahmerichtung, z.B. parallel zu einer Fassade bewegt.

Maßstabsunterschied

Das Ergebnis eines Versuchs ist in Abb. 6.21 dargestellt. Eine vom Boden aus aufgenommene

Bildsequenz, die ein Fachwerkhaus aus kurzer Entfernung abbildet, wurde um acht Bilder erweitert, welche die Szene aus deutlich größerer Distanz zeigt. Die Orientierung der Gesamtsequenz war aufgrund der auftretenden Maßstabsunterschiede nur mit dem neu entwickelten Verfahren zur Bildzuordnung möglich.

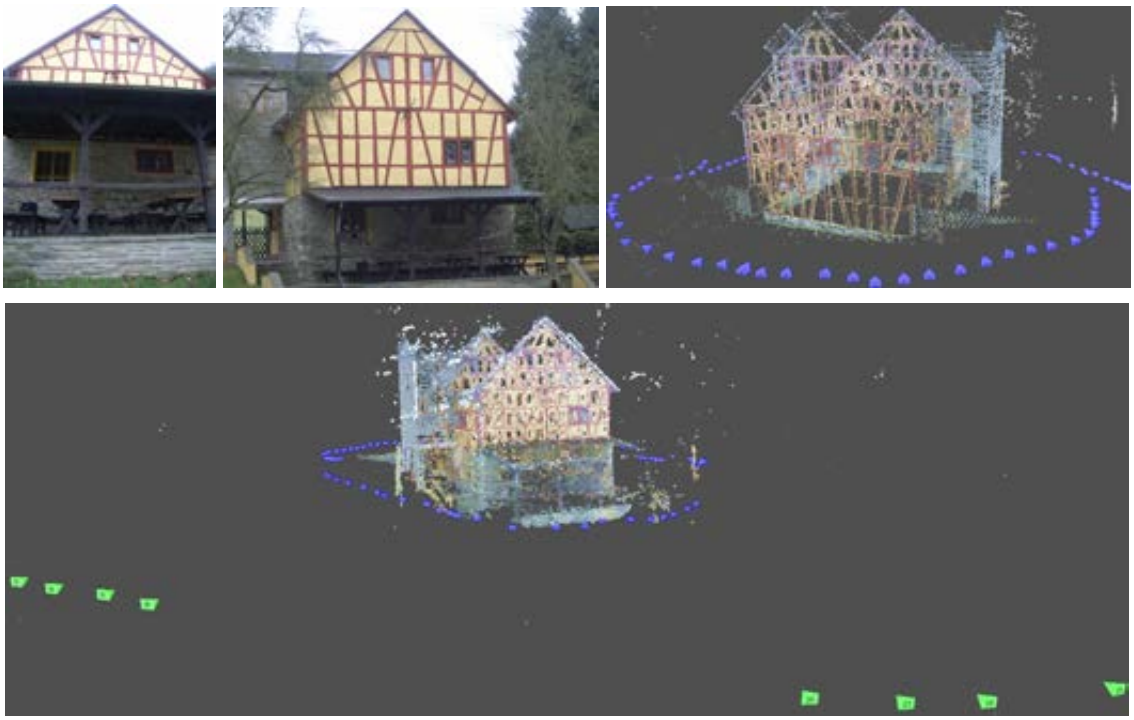


Abbildung 6.21: 3D Rekonstruktion einer Bildsequenz, bei der starke Maßstabsunterschiede auftreten – Bilder (oben links und Mitte) und Visualisierung der 3D Rekonstruktion (oben rechts und unten). 82 Bilder, $\sigma_0 = 0,33$ Pixel.

Divergente Stereopaare

Zur Abbildung großer Szenen kommen zunehmend mobile Erfassungssysteme (Mobile Mapping Systems) zum Einsatz. Diese verwenden typischerweise GPS und inertiale Navigation zur Orientierung der Bildsequenzen. Es konnte demonstriert werden, dass auch derartig gewonnene Bildsequenzen ohne Nutzung der Navigationsinformation direkt orientiert werden können. Eine Sequenz

von niedrig aufgelösten divergenten Stereoinfrarotbildern, die von einem mobilen Erfassungssystem aus aufgenommen wurde, führte zu dem in Abbildung 6.22 dargestellten Ergebnis. Hierbei wurde das Wissen über die Stereokonfiguration nicht genutzt und diese trotzdem mit hoher Genauigkeit rekonstruiert.

Blickwinkelunterschied

In Abschnitt 6.1.2 wurde die Robustheit der Zuordnung gegen Blickwinkeländerung an einem

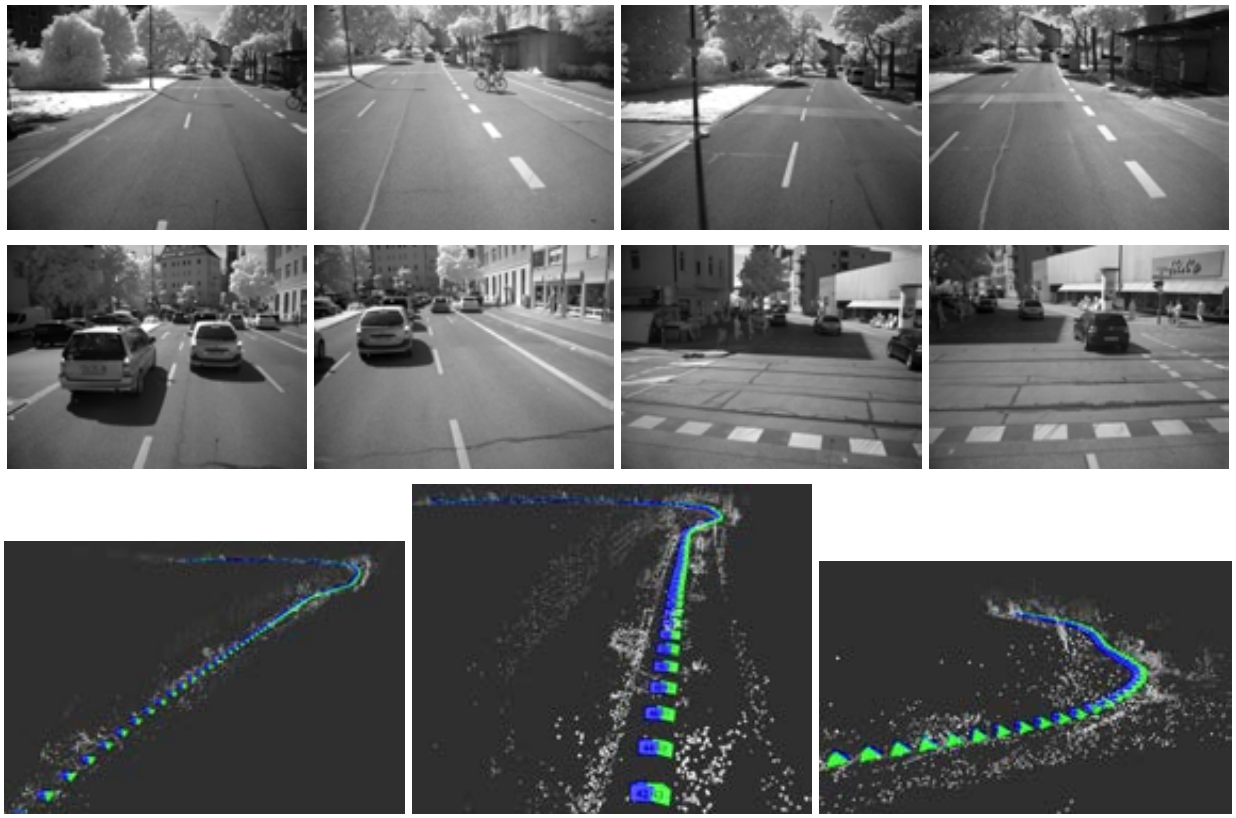


Abbildung 6.22: Bildsequenz mit 2×204 Stereo-Infrarotbildern, welche durch ein mobiles Erfassungssystem aufgenommen wurde. Bildpaare (oben) und Visualisierung der 3D Rekonstruktion mit Punkten und Kameras in Form von Pyramiden (unten, $\sigma_0 = 0.19$ Pixel). Mit freundlicher Unterstützung durch das AGeoBw Dezernat III 1 (1) (Geländeanalyse).

Benchmarkdatensatz evaluiert. Für das Gesamtsystem zur 3D Rekonstruktion wurden weitere praxisnahe Tests durchgeführt. Bei diesen zeigte sich, dass insbesondere durch die Einbeziehung der

geometrischen Plausibilität hinsichtlich Punktdichte und Genauigkeit sehr gute Ergebnisse erzielt werden können. In Abb. 6.23 ist ein Ergebnis für eine Kombination von Boden- und Luftaufnahmen dargestellt. Mit einem kleinen Flugsystem wurde aus ca. 20 Meter Flughöhe mit schräger Blickrichtung der Kamera auf eine Fassade Bilder aufgenommen. Dementsprechend groß ist die Blickwinkeländerung gegenüber den Bodenaufnahmen. Doch auch unter diesen Bedingungen ist in Bereichen mit guter Überdeckung eine genaue Zuordnung möglich, so dass eine größere Bildmenge aus Luft- und Bodenaufnahmen orientiert werden konnte.

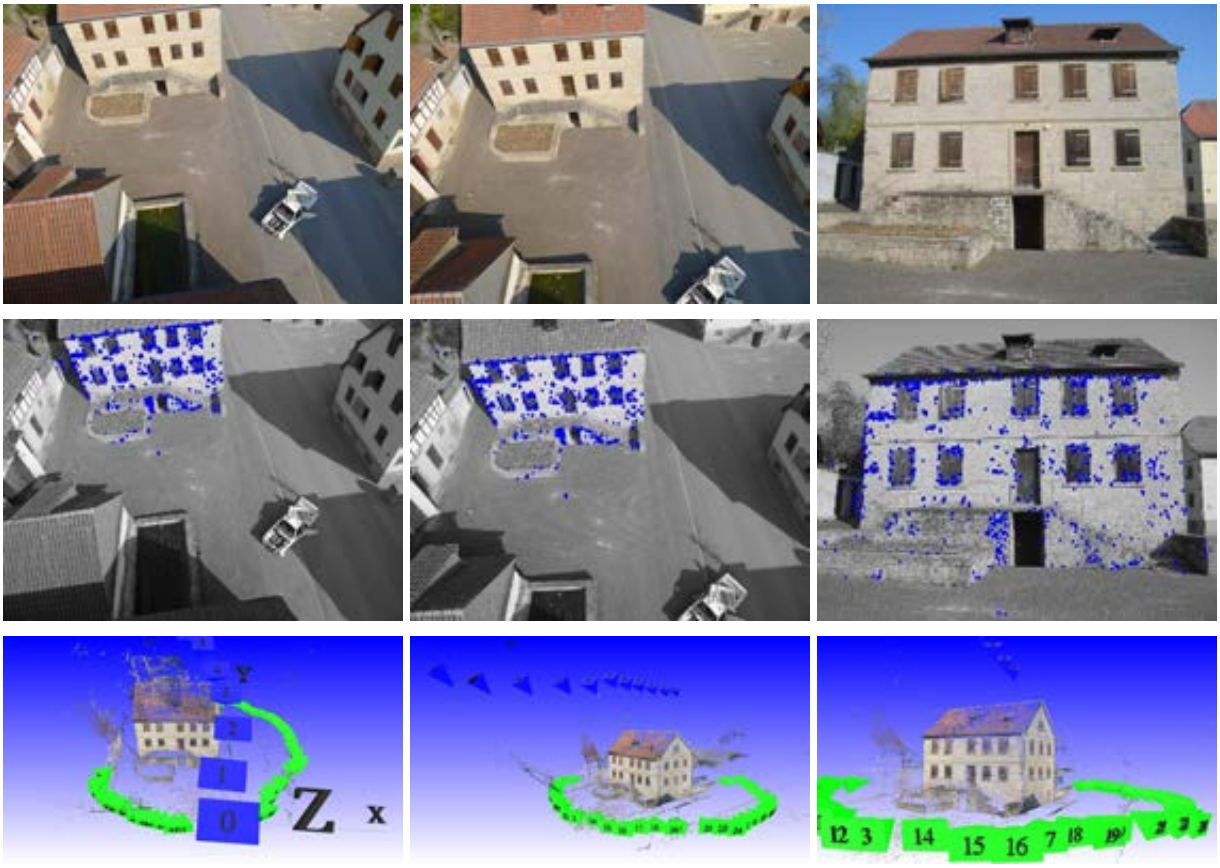


Abbildung 6.23: Ergebnis der Zuordnung von Bildern mit großer Blickwinkeländerung. Oben: Drei Originalbilder mit zwei Luft- und einer Bodenaufnahme; Mitte: 977 Punktzuordnungen; Unten: Relative Orientierung der gesamten Bildmenge aus Luft- und Bodenaufnahmen mit $\sigma_0 = 0,16$ Pixel.

In einem weiteren Versuch wurde eine Sequenz aus sechs Bildern verwendet (siehe Abb. 6.24).

Bei dieser kommen zu einer großen Blickwinkeländerung zwischen den ersten und den letzten drei Bildern die zyklische Struktur der Szene und die schwer zuzuordnende Vegetation. Wird über die große Blickwinkeländerung zugeordnet, so stehen im jeweiligen Bildtripel zwei Kamerapositionen sehr dicht zusammen, während die dritte sehr weit von den übrigen beiden entfernt ist. Dies ist denkbar ungünstig für die Überprüfung der geometrischen Plausibilität, da korrespondierende Punkte aus Bildern der dicht nebeneinander stehenden Kameras zu fast identischen Epipolarlinien im dritten Bild führen. Trotzdem konnten die in Abb. 6.24 unten dargestellten, hinsichtlich 3D-Punktdichte und Genauigkeit sehr guten Ergebnisse erzielt werden.

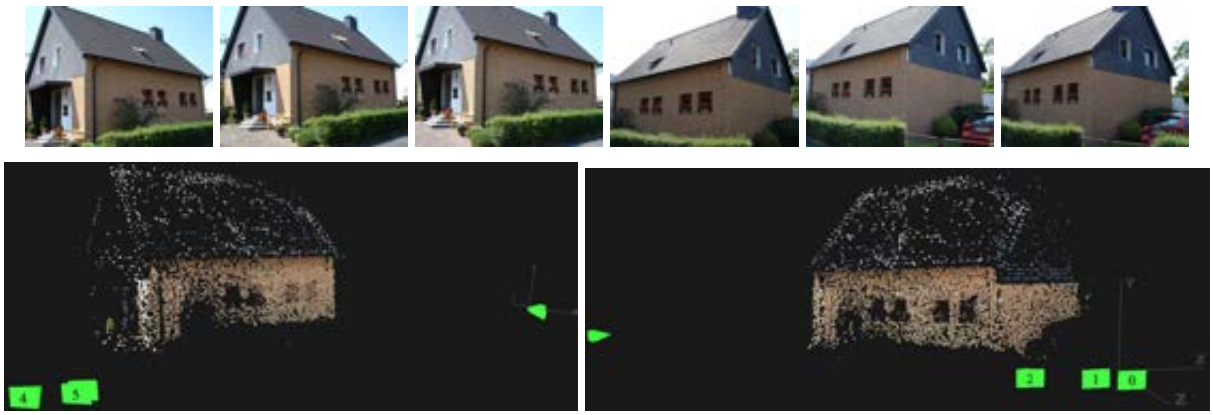


Abbildung 6.24: Ergebnis einer 3D Rekonstruktion einer Szene (unten) auf der Grundlage von sechs Bildern (oben) mit z.T. großer Blickwinkeländerung ($\sigma_0 = 0,10$ Pixel).

Bilder von kleinen unbemannten Flugsystemen

Für einen möglichst umfassenden Eindruck einer urbanen Szene sollten Gebäude möglichst vollständig modelliert werden. Mittels Bodenaufnahmen können nur Schrägdächer rekonstruiert werden und auch für diese sind aufgrund ungünstiger Perspektive meist keine guten Ergebnisse zu erwarten. Grundsätzlich muss daher davon ausgegangen werden, dass für eine detaillierte und vollständige 3D Modellierung eines Gebäudes Luftaufnahmen hinzugezogen werden müssen. Neben herkömmlichen Flugzeugen und automatisierten Fluggeräten (Drohnen), welche ein Gewicht von 200 Kilogramm und deutlich mehr aufweisen, wurde als eine Folge der Fortschritte in der Mikrosystemtechnik die Konstruktion sehr kleiner und leichter Flugsysteme, so genannter „Micro Unmanned Aircraft Systems“ (Micro-UAS) möglich. Ihr Gesamtgewicht beträgt oft lediglich ein

Kilogramm und ihr Durchmesser einen Meter. In Rahmen dieser Arbeit wurden mehrere Versuche mit so genannten „Quadrokoptern“ und ähnlichen UAS gemacht, welche mit vier, sechs oder acht (Quadro-, Hexa- oder Oktokopter) Rotoren nach dem Helikopterprinzip fliegen. Aufgrund dieser Eigenschaften sind Starts und Landungen in bebautem Gebiet und genaues Anfliegen eines Objekts möglich. Quadrokopter können eine Digitalkamera mitführen, was Aufnahmen aus sonst schwer bis gar nicht erreichbaren Perspektiven ermöglicht. Es konnte demonstriert werden, dass derartige Aufnahmen grundsätzlich für Verfahren, geeignet sind, welche auf direkter 3D Rekonstruktion basieren (Mayer & Bartelsen, 2008).

Modellflugzeuge können eine deutlich höhere Nutzlast mitführen als Quadrokopter. Allerdings können Objekte nur aus der Überflugperspektive und mit Bewegung während der Aufnahme aufgenommen werden, Starts und Landungen erfordern freies Gelände. Im Rahmen dieser Arbeit konnte über nahezu unbebautem Gebiet gezeigt werden, dass eine 3D Rekonstruktion mit dem neuen Verfahren möglich ist. Problematisch bei Überflügen über unbebautes, ebenes Gelände ist die schwache Tiefenstruktur. Aufgrund des Verhältnisses zwischen einer typischen Flughöhe von 300 Metern und den geringen Höhenunterschieden am Boden durch Vegetation und Unebenheiten, ist die Bestimmung der Kamerakonstanten nur auf Grundlage der Bilder schwierig bis unmöglich.

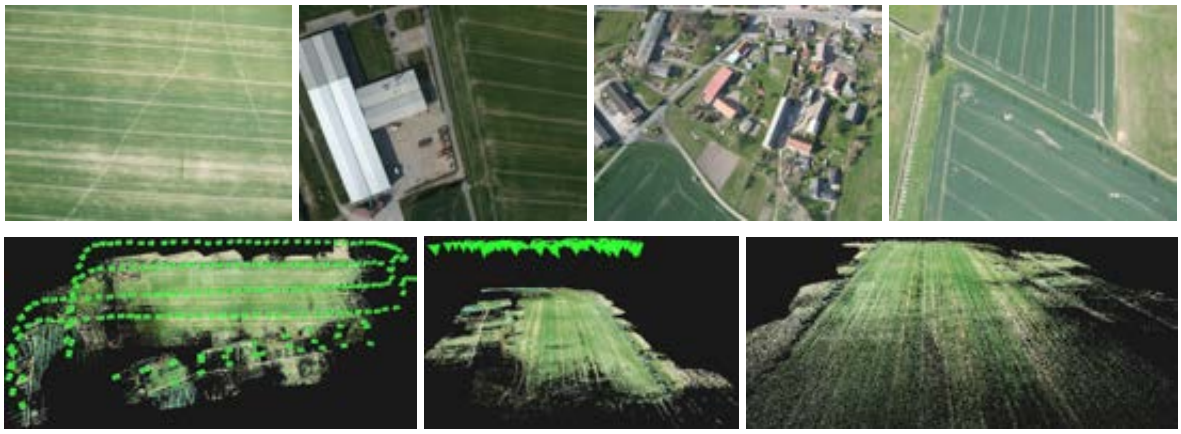


Abbildung 6.25: 3D Rekonstruktion einer in erster Näherung ebenen Szene, die mittels eines Modellflugzeugs mit Autopilotsystem aus ca. 300m Höhe aufgenommen wurde. (Bilder oben und 3D Rekonstruktion unten, 206 Bilder, $\sigma_0 = 0,31$ Pixel.) Mit freundlicher Unterstützung von Dr. rer. nat. Patrick Reidelstürz (HDU Deggendorf).

Daher muss in diesem Fall die Kamerakonstante für die Anwendung genau genug bekannt sein

und kann nicht bei der Bündelausgleichung bestimmt werden. Abb. 6.25 zeigt, dass auf Grundlage der Bilder von Modellflugzeugen großflächige Areale 3D rekonstruiert werden können. Der Detaillierungsgrad von Gebäuden ist aufgrund der Flughöhe jedoch gering. Durch die Verwendung eines Autopilotsystems ist die Aufnahmekonfiguration geordnet und die Flughöhe konstant.

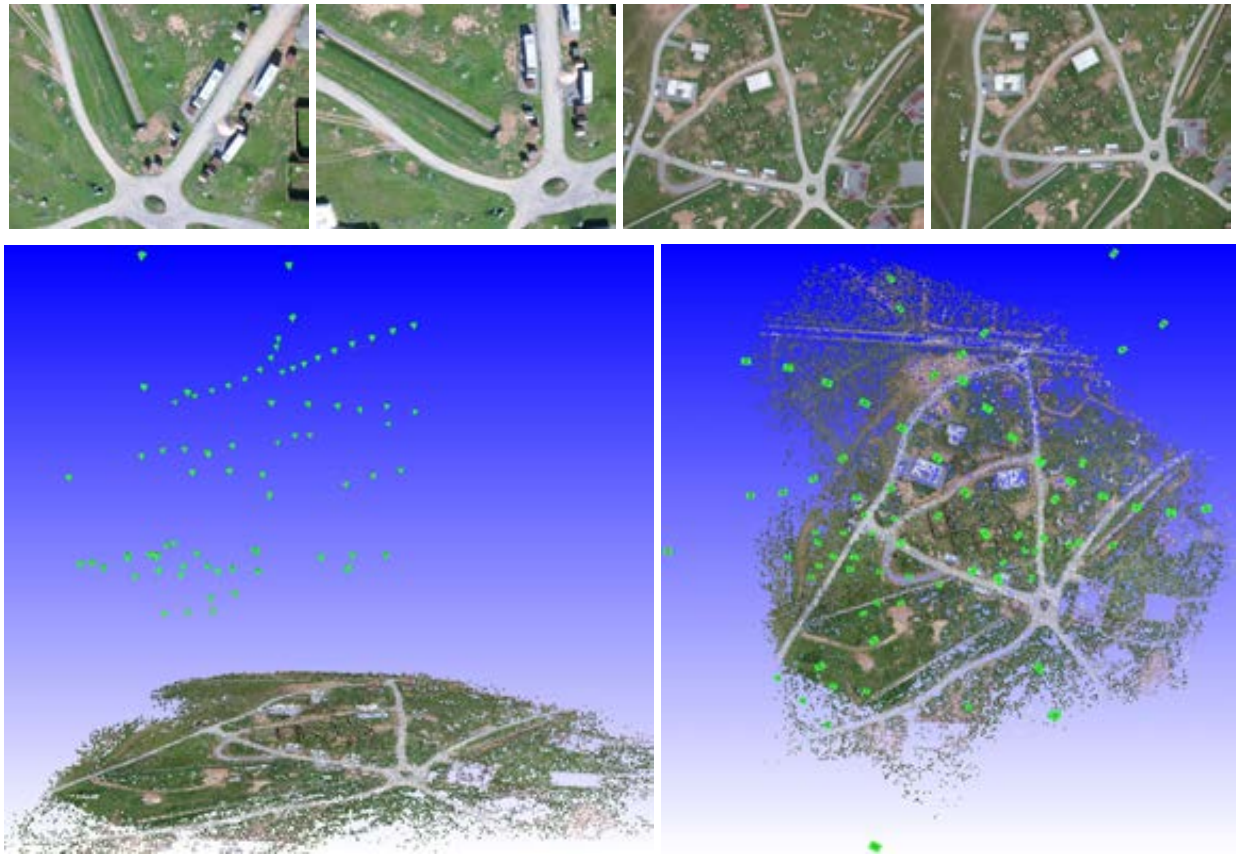


Abbildung 6.26: Relative Orientierung von Luftbildern, welche mittels einer manuellen Befliegung aufgenommen wurden. Dabei treten große Maßstabsunterschiede auf. (Bilder oben und 3D Rekonstruktion unten, 75 Bilder, $\sigma_0 = 0,37$ Pixel.) Mit freundlicher Unterstützung von Dr. rer. nat. Patrick Reidelstürz (HDU Deggendorf).

Doch auch wenn ein System zur automatischen Flugführung nicht zur Verfügung steht, können Bilder von Modellflugzeugen zur 3D Rekonstruktion verwendet werden. Manuelles Befliegen führt aber zu deutlich schwierigeren Aufnahmekonfigurationen, wie Abb. 6.26 zeigt. Erhebliche Maßstabsunterschiede sind genauso zu erwarten, wie große Blickwinkeländerungen und Änderungen

der Lichtverhältnisse. Trotzdem ist mit dem neuen Verfahren eine genaue Relative Orientierung möglich.

Bildsequenzen mit geringen Blickwinkeländerungen zwischen den Aufnahmen begünstigen grundsätzlich die Zuordnung. Bei Videoaufnahmen mit einer Aufnahmefrequenz von 25-30 Bildern pro Sekunde sind geringe Blickwinkeländerungen im Allgemeinen gegeben. Digitale High Definition (HD) Videokameras ermöglichen eine Auflösung, welche eine 3D Rekonstruktion mit hohem Detaillierungsgrad zulässt. Für die in den Abb. 6.27 und 6.28 dargestellten Ergebnisse wurden Videosequenzen verwendet, die von einem Quadrokopter aus aufgenommen wurden. Für die Orientierung der mehrere Minuten langen Videosequenzen wurde jeweils ein Bild pro Sekunde verwendet. Dies führt im gegebenen Fall zu einer sehr hohen Punktdichte. Obgleich eine HD-Video Kamera verwendet wurde, ist die Auflösung mit knapp 1 Megapixel niedrig und die Möglichkeiten zur 3D Rekonstruktion sind dementsprechend begrenzt.

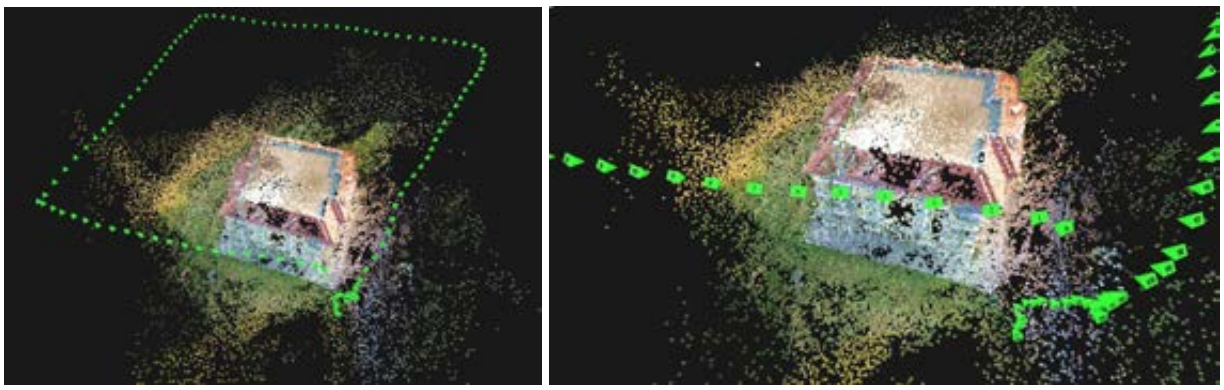


Abbildung 6.27: 3D Punktwolke des Hauses 51 im Übungsdorf Bonnland mit Kamerapositionen von 126 Bildern aus einer Videosequenz mit einer Auflösung von 1.280 x 720 Pixel ($\sigma_0 = 0,25$ Pixel). Mit freundlicher Unterstützung durch das Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung, Ettlingen.

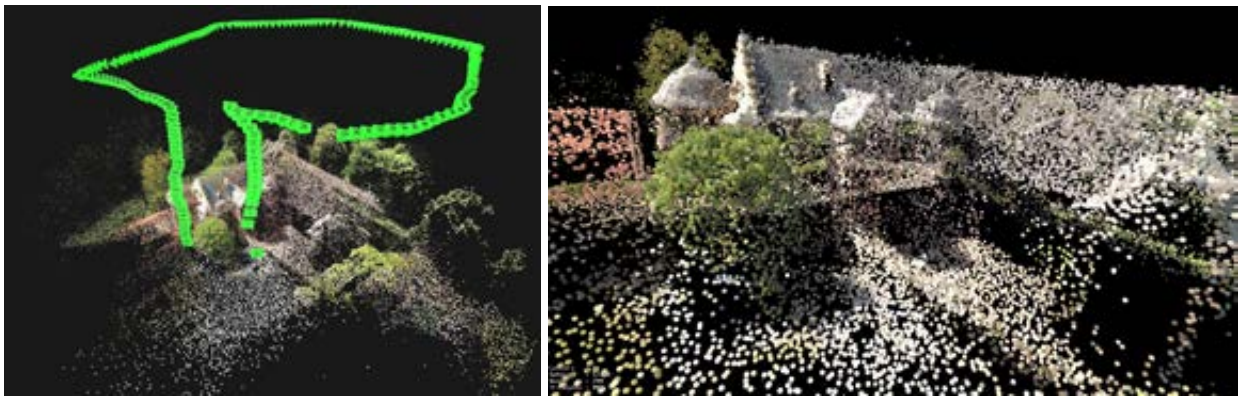


Abbildung 6.28: 3D Punktwolke vom Schloss Greifenstein im Übungsdorf Bonnland und Kamerapositionen von 200 Bildern aus einer Videosequenz mit einer Auflösung von 1.280 x 720 Pixel ($\sigma_0 = 0,19$ Pixel). Mit freundlicher Unterstützung durch das Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung, Ettlingen.

6.3 Bestimmung der Absoluten Orientierung mit Hilfe von GPS Kameras

Zum Test des in Kapitel 5 vorgestellten Ansatzes wurden mit der handelsüblichen GPS Kamera Ricoh Caplio 500SE zahlreiche Bildsequenzen erstellt (Bartelsen & Mayer, 2010a,b). Ziel war es, festzustellen, ob und in wie weit der Ansatz in urbanen Regionen brauchbare Ergebnisse liefern kann und welche Genauigkeiten zu erwarten sind. Eine Kamera mit integriertem GPS Empfänger liefert für jedes Bild eine GPS Koordinate. Werden Bilder einer solchen Kamera relativ orientiert, so ist für die Kamerapositionen eine Zuordnung zwischen Modell- und Weltkoordinaten möglich (siehe Kapitel 5). Bei Verwendung einer handelsüblichen GPS Kamera liegen die Messgenauigkeiten nach Herstellerangaben im Bereich einiger Meter, wobei davon auszugehen ist, dass die Höhenmessung deutlich schlechter ist als die Lagemessung. Probleme werden vorwiegend durch Abschattung durch Gebäude und Vegetation, aber auch durch ungünstige Satellitenkonstellationen verursacht (Börger *et al.*, 2008).

Es wurde zunächst überprüft, ob und in wie weit eine Positionsbestimmung auch in bebautem Gebiet möglich ist. Für die im Rahmen dieser Arbeit verwendete GPS Kamera Ricoh Caplio 500SE sind laut Hersteller bei der Positionsbestimmung Genauigkeiten im Bereich von einem

bis fünf Meter zu erwarten. Zur Überprüfung dieser Angaben wurde ein Testsystem entwickelt. Mit dem Satellitenpositionierungsdienst der deutschen Landesvermessung (SAPOS) ist es möglich, hoch genaue GPS Messungen durchzuführen, deren Abweichungen im Bereich von ein bis drei Zentimetern liegen. Mittels eines speziell angefertigten Stativs wurde die GPS Kamera mit dem SAPOS System so kombiniert, dass das Projektionszentrum der Kamera mit geringem und konstantem Abstand zur SAPOS Antenne fixiert wurde. Auf diese Weise konnte die Position der Kamera im Weltkoordinatensystem sehr genau bestimmt und mit den Positionsbestimmungen des internen Kamera-GPS Systems verglichen werden. Die mittels SAPOS bestimmten Kamerapositionen wurden für alle anschließenden Untersuchungen als Referenzpositionen angesehen. Es wurden ausschließlich jene verwendet, die mit einer Genauigkeit von drei Zentimetern oder weniger bestimmt werden konnten. Der verbleibende Fehler wurde vernachlässigt.

Es wurden mehrere Tests in bebauten Gebieten durchgeführt (siehe z.B. Abb. 6.29 und 6.30) und Statistiken über die Abweichungen der von der GPS Kamera bestimmten Positionen zu den SAPOS Referenzpositionen erstellt. Dabei wurden zunächst die Fehler in Rechtswert, Hochwert und Höhe bestimmt, sowie die daraus resultierende Abweichung in der 2D Position und im Raum. Hierfür wurden Mittelwert und Standardabweichung berechnet (siehe Tab. 6.1).

80 Kamera-Positionen	Fehler Rechtswert	Fehler Hochwert	Fehler Höhe	Fehler Lage (2D)	Fehler Gesamt (3D)
Mittelwert	1,06m	1,91m	4,12m	2,36m	4,97m
Standardabweichung	1,07m	1,72m	3,00m	1,82m	3,17m

Tabelle 6.1: Statistiken über die Abweichungen zwischen den achtzig mittels einer GPS Kamera gemessenen Kamerapositionen und hoch genauen SAPOS Referenzpositionen. Hierbei wurden direkt die von der GPS Kamera gemessenen Höhenwerte verwendet.

Aus Tab. 6.1 ist deutlich erkennbar, dass die Höhenbestimmung durch die GPS Kamera deutlich schlechtere und unzuverlässigere Werte als für die 2D Position liefert, was sich stark auf den Gesamtfehler auswirkt. Unter der Voraussetzung, dass alle Bilder aus etwa derselben Höhe aufgenommen wurden, kann dieses Problem reduziert werden. Die Verwendung des Mittelwertes der Höhenmessung führt zu den in Tab. 6.2 aufgeführten Resultaten. Der unter diesen Bedingungen auftretende systematische Fehler in der Höhenmessung wird vermutlich durch die Geoidkorrektur verursacht. Die Ricoh Caplio 500SE Kamera verfügt über ein internes System zur Geoidkorrektur, welches dem NMEA-0183 Standard entspricht. Ein gleichmäßiger Fehler von drei Metern liegt im Rahmen dessen, was zu erwarten ist.

80 Kamera-Positionen	Fehler Höhe	Fehler Gesamt (3D)
Mittelwert	3,07m	4,07m
Standardabweichung	0,20m	1,32m

Tabelle 6.2: Statistiken über die Abweichungen zwischen achtzig mittels einer GPS Kamera gemessenen Kamerapositionen und hoch genauen SAPOS Referenzpositionen. Hierbei wurden die von der GPS Kamera gemessenen Höhenwerte jeweils durch den Mittelwert ersetzt. Da die Aufnahmen jeweils aus ungefähr gleicher Höhe aufgenommen wurden, reduziert dies die Unzuverlässigkeit der Höhenbestimmung und den resultierenden Gesamtfehler deutlich.

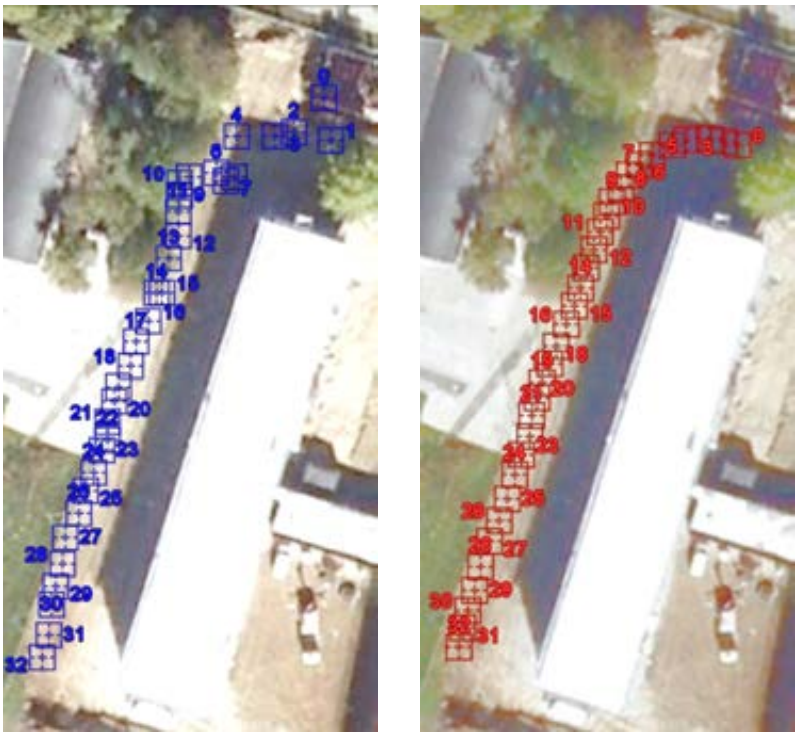


Abbildung 6.29: Vergleich zwischen 33 mit einer GPS Kamera und auf Grundlage des hoch genauen SAPOS Dienstes bestimmten Kamerapositionen. Links: RICOH Kamerapositionen (blau). Rechts: SAPOS Referenzpositionen (rot). Die Kamerapositionen wurden zur Veranschaulichung in Google Earth visualisiert, die mögliche Ungenauigkeit dieses Modells ist dabei unerheblich. Es ist erkennbar, dass die Herstellerangabe plausibel ist. Diese weist eine Genauigkeit von ein bis fünf Meter für die Positionsbestimmung durch die Kamera aus.

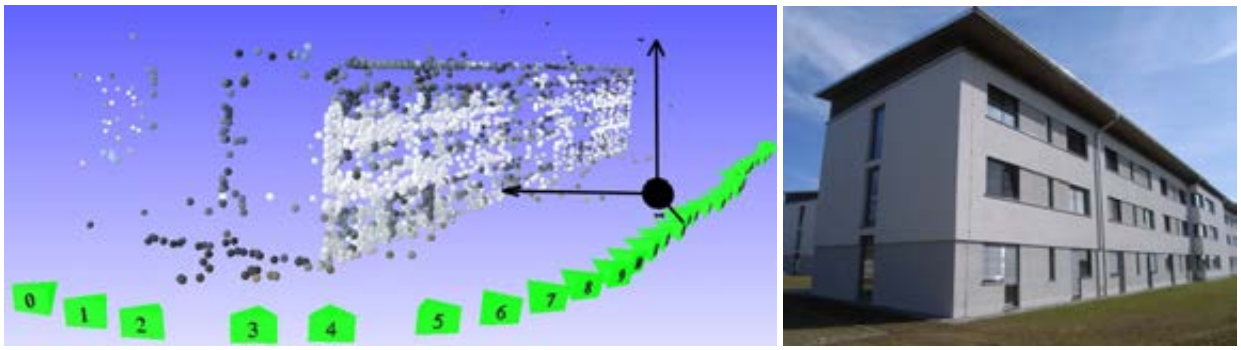


Abbildung 6.30: Links: Kamerapositionen und 3D Punktwolke des Gebäudes 20 der Universität der Bundeswehr München (33 Bilder, $\sigma_0 = 0,32$ Pixel). Die große schwarze Kugel markiert den Schwerpunkt der Kamerapositionen. Rechts: Ein Bild aus der Sequenz.

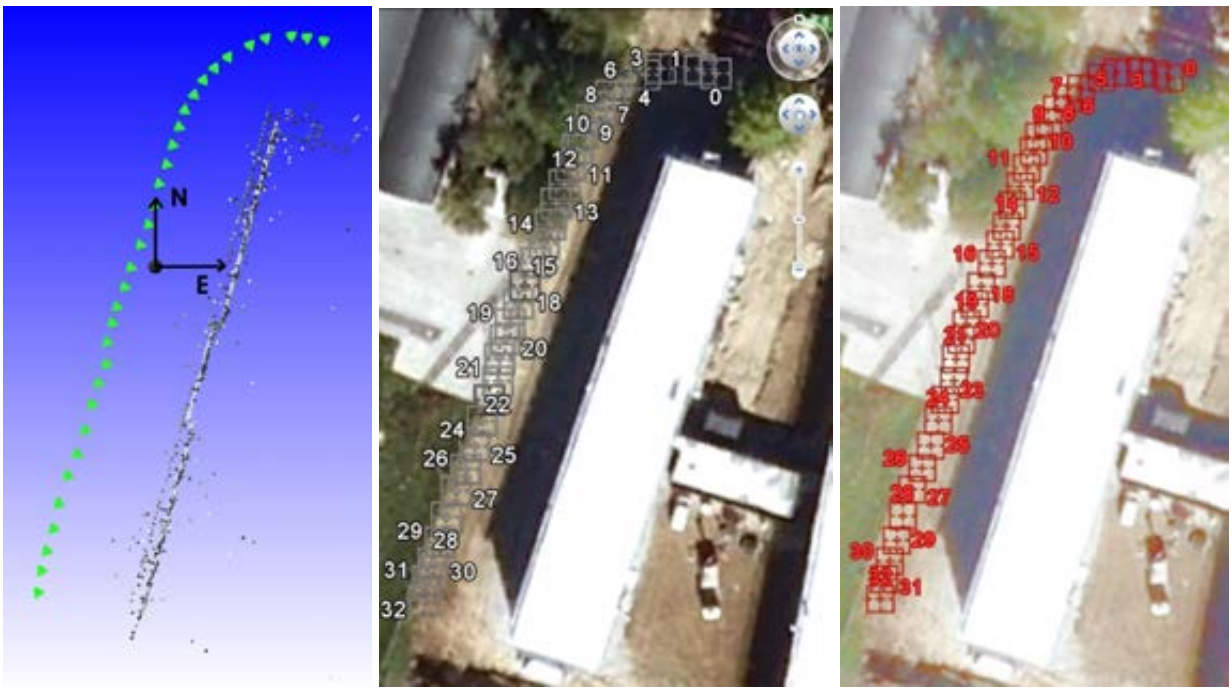


Abbildung 6.31: Links: Kamerapositionen und 3D Punktwolke für das Gebäude 20 der Universität der Bundeswehr München aus der Vogelperspektive. Mitte: Verbesserte Kamerapositionen dargestellt in Google Earth (weiß). Rechts: SAPOS Referenzpositionen (rot).

Im Weitern wurde untersucht, ob und in wie weit die Genauigkeit der Kamerapositionen im Weltkoordinationsystem durch die Verwendung einer genauen Relativen Orientierung in Verbindung mit 3D Ähnlichkeitstransformation und Ausgleichung verbessert werden kann. Dafür wurden die Bilder der in Abb. 6.29 verwendeten GPS Kamera relativ orientiert. Damit ist es möglich, für jede Kameraposition des relativen Modells die Weltkoordinaten zu bestimmen und diese mit den SAPOS Referenzpositionen zu vergleichen. Dieser Vergleich ist in Abb. 6.31 veranschaulicht. Konkret wurde der ursprüngliche Fehler der GPS Messung mit dem Fehler nach der relativen und absoluten Orientierung verglichen, dabei wurde zwischen der Lage (2D) und dem Fehler insgesamt (3D) unterschieden (siehe Tab. 6.3 und 6.4). Die Ergebnisse zeigen, dass eine deutliche Verbesserung erzielt werden konnte.

33 Kamerapos. am Boden (2D)	Durchschnittlicher Abstand von SAPOS	Standardabweichung von SAPOS	Größter Fehler	Kleinster Fehler
GPS Kamera	2,11m	1,67m	8,38m	0,41m
Ausgeglichen	1,36m	0,48m	2,10m	0,70m

Tabelle 6.3: 2D Fehler am Boden – Vergleich Positionsbestimmungen und ausgeglichene Werte der GPS Kamera mit den SAPOS Referenzpositionen.

33 Kamerapos. (3D)	Durchschnittlicher Abstand von SAPOS	Standardabweichung von SAPOS	Größter Fehler	Kleinster Fehler
GPS Kamera	4,32m	2,72m	12,16m	0,42m
Ausgeglichen	3,44m	0,1m	3,69m	3,26m

Tabelle 6.4: 3D Fehler insgesamt – Vergleich Positionsbestimmungen und ausgeglichene Werte der GPS Kamera mit den SAPOS Referenzpositionen.

Eine bestehende Bildsequenz, die relativ orientiert werden konnte, kann (auch nachträglich) mit Bildern einer GPS Kamera kombiniert werden. Dies ist auch dann möglich, wenn die Aufnahmen mit der GPS Kamera zu einem deutlich unterschiedlichen Zeitpunkt aufgenommen wurden. Damit ist dann für die Gesamtsequenz eine absolute Orientierung möglich (Bartelsen & Mayer, 2010a).

6.4 Kombination verschiedener Kameras

Die 3D Rekonstruktion aus Bildern verschiedener Kameras ist aus mehreren Gründen schwieriger als aus Bildern von nur einer Kamera. Sowohl für die Bestimmung der Epipolargeometrie mittels des 5-Punkt-Algorithmus (siehe Abschnitt 2.1) als auch für die Bündelausgleichung (siehe Abb. 2.3) sind die verschiedenen Kamerakalibrierungen zu berücksichtigen.

Um Gebäude möglichst vollständig und detailliert modellieren zu können, sind Bilder von Fassaden und Dächern aus jeweils günstiger Perspektive notwendig. Hierfür eignet sich die Kombination von Luft- und terrestrischen Bildern. Allerdings muss bei dieser Konfiguration grundsätzlich davon ausgegangen werden, dass große Blickpunktänderungen auftreten.

Durch verschiedene Bildgrößen und Auflösungen wird eine Szene im Allgemeinen bei der Verwendung verschiedener Kameras mit verschiedenen Maßstäben abgebildet. Dies führt zu globalen Maßstabsunterschieden zwischen den Bildern einer Szene, auch z.B., wenn diese vom gleichen Punkt aus aufgenommen wurden. Da nicht vorausgesetzt werden kann, dass die Aufnahmekonfigurationen verschiedener Kameras aufeinander abgestimmt sind, müssen große Blickpunkt- und -winkeländerungen mit einbezogen werden. Große Blickpunkt- und -winkeländerungen führen zu lokalen Maßstabsvarianzen und perspektiver Verzerrung. Zudem können die Lichtverhältnisse stark unterschiedlich sein. Wie die in den Abbildungen 6.21 und 6.32 dargestellten Ergebnisse zeigen, können mit dem neuen Ansatz zur robusten Bildzuordnung auch auf Grundlage deutlich verschiedener Kameramodelle sehr gute Ergebnisse erzielt werden. Dabei wurden folgende Aspekte getestet:

1. Bildzuordnung und 3D Rekonstruktion bei zwei verschiedenen Kameras mit deutlich unterschiedlichen Bildgrößen, Bildseitenverhältnissen und Kamerakalibrierungen: Um stark verschiedene Bildseitenverhältnisse zu erhalten, wurden die Bilder einer Kamera um 90° gedreht.

2. Bildzuordnung und 3D Rekonstruktion mit mehreren verschiedenen Kameras und erheblichen Maßstabsunterschieden: Um eine relative Orientierung von derartigen Aufnahmekonfigurationen durchführen zu können, ist ein Zuordnungsverfahrens erforderlich, welches eine hohe Robustheit bzgl. globaler und lokaler Maßstabsvarianzen, perspektivischer Verzerrung und Änderung der Lichtverhältnisse aufweist.

Die derzeit gängigen Verfahren zur Bildzuordnung sind jedoch nach (Morel & Yu, 2009) in mindestens einem dieser Punkte stark eingeschränkt. Auch zu diesem Zweck wurde der in Kapitel 4 vorgestellte Ansatz entwickelt. Die Ergebnisse in den Abb. 6.21 und 6.32 zeigen, dass er robust hinsichtlich der erläuterten Probleme ist.

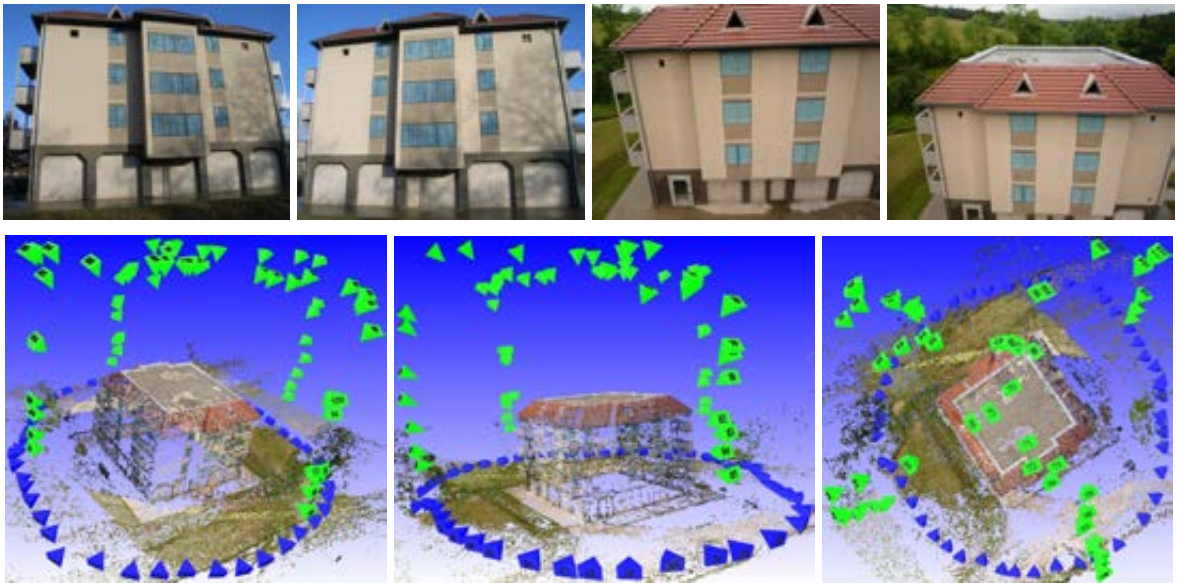


Abbildung 6.32: Haus 51 Bonnland, Kombination von Boden- und Luftaufnahmen mit zwei verschiedenen Kameras (112 Bilder, $\sigma_0 = 0.26$ Pixel).

Kapitel 7

Bewertung der Ergebnisse

7.1 Robustheit des neu entwickelten Ansatzes

Die Ergebnisse in Kapitel 6 zeigen, dass der neu entwickelte Ansatz eine hohe Robustheit gegenüber Maßstabsunterschieden und Änderungen der Lichtverhältnisse aufweist. Des Weiteren wurden bei der Änderung des Blickwinkels Ergebnisse erzielt, die nur von ASIFT übertroffen werden. Bezüglich Unschärfe und JPEG-Komprimierung konnte ebenfalls eine hohe Robustheit nachgewiesen werden. Der Testbildsatz unterscheidet bei den Tests auf Unschärfe, Blickwinkeländerung und Rotation um die Hauptachse und Maßstabsänderung jeweils zwischen Szenen mit starker und schwacher Textur. Es ist erkennbar, dass der neu entwickelte Ansatz für stark texturierte Szenen bessere Ergebnisse liefert. Derartige Szenen sind typisch für urbane Gebiete, welche viele Altbauten aufweisen.

Die in Kapitel 6 für den neuen Ansatz bzgl. der Robustheit gegenüber Maßstabsunterschieden dargestellten Ergebnisse sind allgemein sehr gut und hinsichtlich der Anzahl der korrekten Zuordnungen deutlich besser als SIFT und ASIFT. Dies ist vor allem dadurch zu erklären, dass bei der Bildzuordnung Bildausschnitte verwendet werden, die weniger stark geglättet werden. Dadurch bleiben die Bildausschnitte stärker unterscheidbar, so dass mehr korrekte Zuordnungen erzielt werden. Dies ist ein entscheidender Unterschied zur SIFT-Zuordnung. SIFT-Deskriptoren werden stets auf der Gaußebene bestimmt, für die ein Maßstabsraumextremwert gefunden wurde, wodurch die Bildinformation in der Regel deutlich stärker geglättet wird.

Lowe (2004) führte Tests zur Bestimmung der optimalen Anzahl der Maßstabsraumebenen pro Oktave bei der Bestimmung von SIFT-Punkten durch. Diese Tests zielten auf die Optimierung der SIFT-Zuordnung ab und führten zu dem Ergebnis, dass die Verwendung von mehr als

drei Maßstabsraumebenen keinen Nutzen bringt. Lowe (2004) vermutete aber, dass mehr Ebenen für andere Anwendungen durchaus sinnvoll sein könnten. Die Ergebnisse dieser Arbeit werden somit als Bestätigung dieser Annahme gewertet. Konkret konnte gezeigt werden, dass die zusätzlichen SIFT-Punkte für die normierte Kreuzkorrelation und die Affine Kleinste Quadrate Zuordnung verwendet werden können. Bei der Generierung der SIFT-Punkte wurden hier zehn Maßstabsraumebenen verwendet und die Schwellwerte für Eckenstrukturen und Gaußdifferenzen deutlich angepasst, was durch die Verwendung der leistungsstarken GPU-Implementierung von Wu (2007) ohne größere Laufzeitverluste für das Gesamtsystem umsetzbar war.

Die deutlich höhere Anzahl an skalierten Punkten führt durch die schwächere Glättung, die Verwendung von normierter Kreuzkorrelation und Affiner Kleinster-Quadrate-Zuordnung sowie die Einbeziehung der geometrischen Plausibilität zu einer deutlich robusteren Zuordnung bei Maßstabsunterschieden und auch bei Blickwinkeländerungen. Auf der Grundlage des neuen Ansatzes ist die Orientierung von Bildsequenzen mit großen Maßstabsunterschieden, an denen die SIFT Zuordnungsmethode scheitert, zuverlässig und mit hoher Genauigkeit möglich. Dies belegen die Tests des Gesamtsystems an den eigenen, praxisnahen Bildsequenzen. Die Ursache dafür liegt in der höheren Anzahl an korrekten Zuordnungen. SIFT liefert auch bei starken Maßstabsunterschieden nur eine kleinere Zahl an korrekten Zuordnungen, wenn auch mit einem sehr geringen Fehleranteil. Für eine robuste Parameterschätzung ist die Zahl der Zuordnungen und deren Genauigkeit jedoch bereits bei der Zuordnung von zwei Bildern zu gering. Aufgrund der Ergebnisse an den Testbilddatensätzen werden ähnliche Probleme für die ASIFT Methode erwartet. Letztere konnte im Rahmen dieser Arbeit jedoch nicht in ein Gesamtsystem integriert werden.

7.2 Genauigkeit des neu entwickelten Ansatzes

Neben insgesamt der hohen Anzahl an Zuordnungen konnte für alle Testkriterien jeweils eine hohe Genauigkeit nachgewiesen werden. Diese ist eine notwendige Voraussetzung zur Verwendung der relativen Orientierung für Verfahren zur dichten Tiefenschätzung. Bereits ein durchschnittlicher Rückprojektionsfehler (σ_0) zwischen ein und zwei Pixeln führt z.B. für die Semiglobale Zuordnung (Hirschmüller, 2008) in der Regel dazu, dass keine qualitativ hochwertigen Tiefenbilder bestimmt werden können. Die absolute Anzahl und der Anteil an Punktkorrespondenzen, die z.T. mit hoher Subpixel-Genauigkeit bestimmt werden konnten, sind bei dem Ansatz dieser Arbeit durchgehend sehr hoch, zum überwiegenden Teil höher als bei allen übrigen getesteten Verfahren. Das Gesamtsystem liefert auch für schwierige Aufnahmekonfigurationen mit einer größeren Zahl

an Bildern Ergebnisse, deren durchschnittlicher Rückprojektionsfehler in der Regel zwischen 0,2 bis 0,3 Pixel liegt (Bartelsen *et al.*, 2012a,b).

7.3 Grenzen des Ansatzes

Die Ergebnisse für den Testbilddatensatz weisen durchgehend eine hohe Zahl an korrekten Zuordnungen auf. Auch wenn der Anteil der korrekten Zuordnungen zumeist niedriger als für die auf der SIFT Zuordnungsmethode basierenden Ansätze ist, so ist er doch vergleichbar. Dies entspricht nicht dem, was vor dem Beginn dieser Arbeit erwartet wurde, nämlich, dass die Einbeziehung der geometrischen Plausibilität bereits bei der Zuordnung von zwei Bildern zu einem deutlich höheren Anteil an korrekten Zuordnungen führt. Die vorgestellten Ergebnisse sprechen demzufolge für die Überprüfung der Ergebnisse durch Hinzunahme eines dritten Bildes.

Der Rechenaufwand ist durch die hohe Anzahl an Punkten, der großen Menge an verwendeter Bildinformation und die Bündelausgleichung hoch. Die Orientierung von einhundert hoch aufgelösten Bildern benötigt bereits mehrere Stunden Rechenzeit. Allerdings erscheint eine weitere Optimierung bzgl. der Laufzeit möglich. Der Aufwand an Hauptspeicher ist ebenfalls hoch. Die Ursache dafür liegt in der Verwendung zahlreicher Gaußebenen in Verbindung mit vielen Bildausschnitten und den zugehörigen Intensitätswerten. Für ein Rechensystem sollten je nach Bildauflösungen und Zahl der SIFT-Punkte 8 bis 16 Gigabyte Hauptspeicher vorgesehen werden.

Die in Abschnitt 4.3 beschriebene Vorgehensweise zur Bildzuordnung sieht für jedes Punktepaar individuell die Auswahl der geeigneten Bildauflösung vor. Damit ist eine effiziente Umsetzung zu einer GPU-Implementierung kaum möglich, da für dessen Laufzeit die Speicherverwaltung der Programmdaten von ausschlaggebender Bedeutung sind.

Bislang wird die Affine Kleinste-Quadrate-Zuordnung als lokale Optimierung nach der normierten Kreuzkorrelation angewandt. Die normierte Kreuzkorrelation für sich ist wenig robust gegen perspektive Verzerrung, folglich werden korrekte Zuordnungen bei starken Blickwinkeländerungen früh verworfen. Der in dieser Arbeit entwickelte Ansatz kann dementsprechend mit der Simulation der Blickwinkeländerungen vor bzw. in Verbindung mit der Korrelation (siehe Abschnitt 8.2) verbessert werden, da dieses Konzept das größte Potential für eine robuste Bildzuordnung bei starken Blickwinkeländerungen hat.

7.4 Praxistauglichkeit

In dieser Arbeit konnte demonstriert werden, dass der entwickelte Ansatz geeignet ist, um damit Bildblöcke zu orientieren, welche mit geringem technischen Aufwand, z.B. mit einfachen Konsumerkameras, erstellt worden sind. Anders als bei vielen anderen Ansätzen, können auch Bilder mit großer Basis orientiert werden. Die relative Orientierung kann mit Subpixel-Genauigkeit ohne Vororientierung durch INS-Systeme bestimmt werden, so dass genaue Tiefenbilder generiert werden können. Als Datenquelle können von Hand oder von kleinen, leichten, verhältnismäßig billigen Trägersystemen aus aufgenommene Aufnahmen verwendet werden. Auch wenn der Rechen- und Speicheraufwand hoch ist, kann Standard-PC Hardware genutzt werden. Die Verwendung durch den Nutzer erfordert wenig Expertenwissen und erfolgt kommandozeilenbasiert. Für eine kommerzielle Nutzung erscheint jedoch ein kompakter Workflow unverzichtbar, welcher alle Prozessierungsschritte ausgehend von den Bilddaten bis hin zum standardisierten 3D Modell umfasst.

7.5 Fazit

Die Ergebnisse dieser Arbeit zeigen neue Möglichkeiten der Nutzung von Bildern von einfachen Kameras zur 3D-Modellierung auf. Die Fähigkeit, genaue Orientierungen für quasi beliebige Bildmengen bestimmen zu können, ermöglicht die Anwendung von weiteren Methoden aus Photogrammetrie und Computer Vision zur dichten Tiefenschätzung, Oberflächenrekonstruktion und Objektextraktion. Für eine ausreichende Genauigkeit ist die Verwendung von technisch aufwändigen und teuren Intertial- oder Differential-GPS Systemen nicht zwingend erforderlich, da diese Arbeit belegt, dass diese oft auch allein über Bildzuordnungsverfahren erzielt werden kann. Detaillierte 3D Modelle können somit zu geringen Kosten generiert und damit von einen sehr großen Nutzerkreis verwendet werden.

Kapitel 8

Zusammenfassung und Ausblick

8.1 Zusammenfassung

Mit dieser Arbeit wurde eine Bildzuordnungsverfahren präsentiert, welches gegen Maßstabsunterschiede und Änderung der Lichtverhältnisse deutlich robuster ist, als bisherige Verfahren. Auch gegen andere Einflüsse, die bei einer praxisnahen Anwendung erwartet werden müssen, konnte ein hohes Maß an Robustheit nachgewiesen werden. Ziel war es, die relative Orientierung von Bildern mit schwieriger Aufnahmekonfiguration zu ermöglichen. Mit der Integration in ein Gesamtsystem können große Bildblöcke orientiert werden. Es wurde gezeigt, dass auf Grundlage genauer relativer Orientierung und GPS-Information von integrierten Empfängern in vielen Fällen eine absolute Orientierung mit zufriedenstellender Genauigkeit möglich ist.

Objektextraktion, dichte Tiefenschätzung und Oberflächenrekonstruktion setzen häufig die Kenntnis der relativen Orientierung voraus. Dabei ist auch deren Genauigkeit ein wichtiger Aspekt. Für die relative Orientierung ist Information über die Beschaffenheit einer Szene zwar hilfreich, aber keine notwendige Voraussetzung. Wie in den Kapiteln 2 und 3 erläutert wurde, kann diese oft allein auf Grundlage und Punktkorrespondenzen zwischen den Bildern bestimmt werden. Damit kommt der Bildzuordnung eine große Bedeutung zu.

8.2 Ausblick

Die bisherige Umsetzung verwendet die SIFT-GPU Implementierung von Wu (2007) zur Extraktion markanter Punkte inklusive deren Skalierung. Der SFOP-Operator (Förstner *et al.*, 2009) bietet dieselbe Möglichkeit, liefert jedoch z.T. deutlich mehr stabile Punkte für schwächer texturierte

Bildregionen, was ihn für die Anwendung zur Rekonstruktion von urbanen Bereichen geeigneter erscheinen lässt. Die in dieser Arbeit durchgeführten Benchmarktests sollten daher auf der Grundlage von SFOP-Punkten wiederholt werden.

Bildzuordnung bei großen Blickwinkeländerungen sollte über die Möglichkeiten des in Abschnitt 3.3 erläuterten Ansatzes ASIFT hinaus noch deutlich verbessert werden können.

Abbildung 8.1: Verwendung einer projektiven Transformation zur Angleichung perspektiver Verzerrung. Oben: Originalbilder. Unten links: Transformiertes Bild (Simulierte 60° Drehung) und mittels FASIAM gefundene 1.043 Punktkorrespondenzen (gelb). Unten rechts: Punktkorrespondenzen im zweiten Bild (unverändert).

Aufgrund der fehlenden geeigneten Normierung bezüglich perspektiver Verzerrung für einen Punktoperator erscheint die Simulation der Blickwinkeländerung als beste Methode. Als Modell zur Simulation sollte jedoch nicht die affine, sondern die projektive Transformation verwendet werden. Diese hat den Vorteil, dass pro Ebene nur eine Transformation erforderlich ist, was die Zahl

der benötigten Transformationen für den häufigen Fall von Ebenen im Raum deutlich reduzieren würde (siehe Abb. 8.1). Liu *et al.* (2012) zeigen, dass diese Vorgehensweise zu besseren Ergebnissen führt. Affine wie projektive Transformation sind Standardoperationen von GPUs, so dass gute Voraussetzungen für eine performante Umsetzung dieser Konzepte gegeben sind.

In (Schaffalitzky & Zisserman, 2002) und (Yao & Cham, 2007) wurden Methoden zur Ordnung, d.h. Bestimmung der Überlappung der Bilder von Bildmengen präsentiert. Letztere wurden für den neu entwickelten Ansatz für eine prototypische Realisierung zur automatischen Bestimmung von Bildblöcken verwendet. Ein wichtiger Unterschied zu anderen Ansätzen, wie z.B. Bundler (siehe Abschnitt 3.4), ist die Verwendung von Bildtripeln, was zu einer deutlich erhöhten Zuverlässigkeit führt. Eines der ersten Ergebnisse ist in Abbildung 8.2 dargestellt.

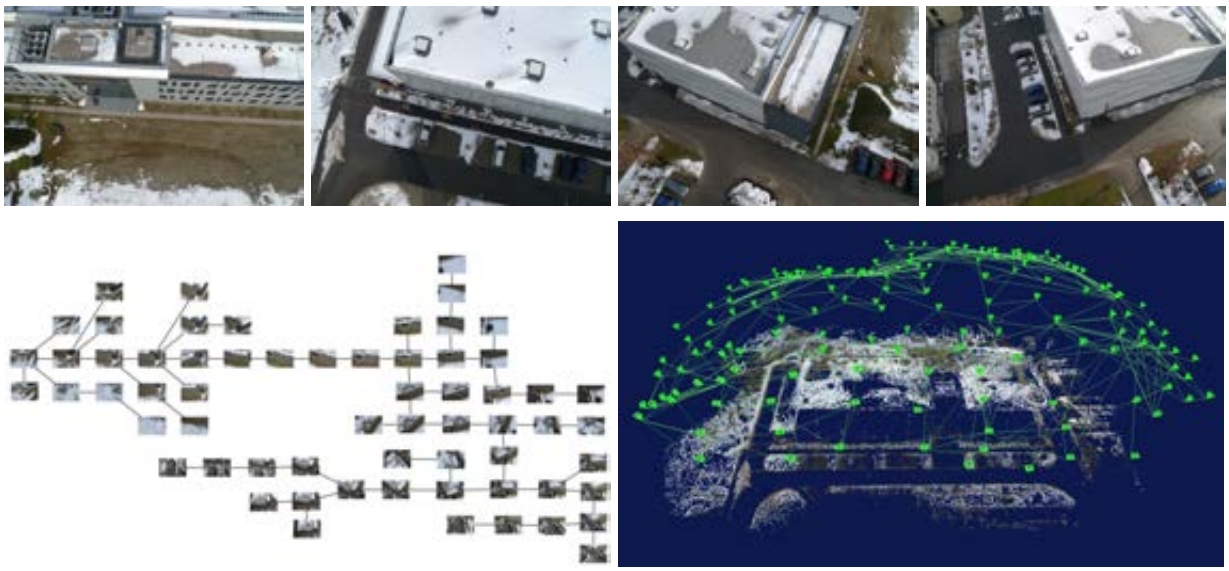


Abbildung 8.2: Ergebnis eines Prototypen zur automatischen Auffindung von Überlappungen für den Drei-Bild-Fall. Oben: Ungeordnete Bilder (Institut für Robotik und Mechatronik, DLR) eines Gebäudes. Unten links: Automatisch generierter Spannbaum. Unten rechts: Ergebnis der relativen Orientierung mit Kennzeichnung der verknüpften Bildpaare.

Aufgrund der hohen Genauigkeiten des entwickelten Ansatzes sind die damit bestimmten relativen und absoluten Orientierungen gut als Grundlage für Verfahren zur dichten Tiefenschätzung geeignet (Mayer *et al.*, 2012). Viele bisherige Arbeiten aus Computer Vision verwenden Bildsequenzen mit kleiner Basis und geringer Bildauflösung. Für hoch aufgelöste Bilder mit großer

Abbildung 8.3: Dichte 3D Punktwolke – Ergebnis dichte Tiefenschätzung mittels Semiglobal Matching – SGM (Hirschmüller, 2008) auf der Grundlage einer durch FASIAM bestimmten hoch genauen relativen Orientierung.

Basis müssen Probleme wie perspektive Verzerrung und Verdeckung angegangen werden. Am Institut für Robotik und Mechatronik des Deutschen Zentrums für Luft- und Raumfahrt (DLR) wurde von Hirschmüller (2008) ein leistungsstarker Ansatz zur dichte Tiefenschätzung entwickelt, welcher aufgrund der hohen Genauigkeit der relativen Orientierung die Ergebnisse des neuen Ansatzes als Grundlage verwenden kann (siehe Abb. 8.3). In den kommenden Jahren soll untersucht werden, wie weit auf der Grundlage beider Ansätze eine detaillierte 3D dichte Tiefenschätzung sowie 2,5D und voll 3D Rekonstruktion für Bilder mit großer Basis möglich ist.

Literaturverzeichnis

AGARWAL, S., SNAVELY, N., SIMON, I., SEITZ, S.M. & SZELISKI, R. (2009). Building Rome in a Day. In *12th IEEE International Conference on Computer Vision (ICCV'09)*. 25, 26

BARTELSEN, J. & MAYER, H. (2010a). Orientation of Image Sequences Acquired from UAVs and with GPS Cameras. *Surveying and Land Information Sciences*, **70**, 151–159. 44, 83, 87

BARTELSEN, J. & MAYER, H. (2010b). Orientation of Image Sequences Acquired from UAVs and with GPS Cameras. In *Proceedings of European Calibration and Orientation Workshop (EUROCOW)*. 83

BARTELSEN, J., MAYER, H., HIRSCHMÜLLER, H., KUHN, A. & MICHELINI, M. (2012a). Orientation and Dense Reconstruction from Unordered Wide Baseline Image Sets. *PFG Photogrammetrie, Fernerkundung, Geoinformation*, **2012**, 421–432. 92

BARTELSEN, J., MAYER, H., HIRSCHMÜLLER, H., KUHN, A. & MICHELINI, M. (2012b). Orientation and Dense Reconstruction of Unordered Terrestrial and Aerial Wide Baseline Image Sets. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, **I-3**, 25–30. 92

BIGÜN, J. (1990). A Structure Feature for Some Image Processing Applications Based on Spiral Functions. *Computer Vision, Graphics, and Image Processing*, **51**, 166–194. 19, 29

BÖRGER, K., GASPER, S., LICKFETT, B. & TOURNAY, K. (2008). Auswirkung von Störungen auf die Navigation mit GPS. *Allgemeine Vermessungs-Nachrichten, Ausgabe 10/2008*, 338–345. 83

FISCHLER, M. & BOLLES, R. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, **24**, 381–395. 13, 54

FÖRSTNER, W. (1994). A Framework for Low Level Feature Extraction. In *Third European Conference on Computer Vision*, vol. II, 383–394. 16

FÖRSTNER, W. & GÜLCH, E. (1987). A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In *ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, 281–305. 11, 16

FÖRSTNER, W., DICKSCHEID, T. & SCHINDLER, F. (2009). Detecting interpretable and accurate scale-invariant keypoints. In *12th IEEE International Conference on Computer Vision (ICCV'09)*, 2256–2263. 19, 29, 35, 94

FRAHM, J.M., FITE-GEORGEL, P., GALLUP, D., JOHNSON, T., RAGURAM, R., WU, C., JEN, Y.H., DUNN, E., CLIPP, B., LAZEBNIK, S. & POLLEFEYS, M. (2010). Building Rome on a Cloudless Day. In *11th European conference on Computer vision: Part IV*, ECCV'10, 368–381. 27

GOOGLE (2011). Google Earth Homepage. `http://www.google.de/intl/de/earth/index.html`. 3

GRÜN, A. (1985). Adaptive Least Squares Correlation: A Powerful Image Matching Technique. *South African Journal of Photogrammetry, Remote Sensing and Cartography*, **14**, 175–187. 23

HARRIS, C. & STEPHENS, M. (1988). A Combined Corner and Edge Detector. In *Alvey Conference*, 147–152. 11, 16, 57

HARTLEY, R. & ZISSERMAN, A. (2004). *Multiple View Geometry in Computer Vision – Second Edition*. Cambridge University Press, Cambridge, UK. 2, 7, 11, 45

HIRSCHMÜLLER, H. (2008). Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**, 328–341. 91, 97

KOCH, R., POLLEFEYS, M. & VAN GOOL, L. (1999). Robust Calibration and 3D Geometric Modeling from Large Collections of Uncalibrated Images. In *Mustererkennung 1999*, 413–420. viii, 1, 7

KOENDERINK, J. (1984). The Structure of Images. *Biological Cybernetics*, **50**, 363–370. 15

LI, X., WU, C., ZACH, C., LAZEBNIK, S. & FRAHM, J.M. (2008). Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs. In *10th European Conference on Computer Vision*, ECCV'08, 427–440. 2, 27

LINDEBERG, T. (1994). *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Boston, USA. 17, 24, 37

LINDEBERG, T. (1998). Feature Detection with Automatic Scale Selection. *International Journal of Computer Vision*, **30**, 79–116. 24, 38

LIU, W., WANG, Y., CHEN, J., GUO, J. & LU, Y. (2012). A Completely Affine Invariant Image-Matching Method Based on Perspective Projection. *Machine Vision and Applications*, **23**, 231–242. 96

LOURAKIS, M.A. & ARGYROS, A. (2009). SBA: A Software Package for Generic Sparse Bundle Adjustment. *Association for Computing Machinery Transactions on Mathematic Software*, **36**, 1–30. 26

LOWE, D. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, **60**, 91–110. 11, 16, 18, 19, 20, 32, 37, 38, 56, 90, 91

LOWE, D. (2005). Demo Software: SIFT Keypoint Detector. `http://www.cs.ubc.ca/~lowe/keypoints/`. 56

MATAS, J., CHUM, O., URBAN, M. & PAJDLA, T. (2002). Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *British Machine Vision Conference*, 384–393. 24, 57

MAYER, H. (2002). Estimation of and View Synthesis with the Trifocal Tensor. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. (34) 3A, 211–217. 44

MAYER, H. (2007a). Automatische Orientierung mit und ohne Messmarken. Das Mögliche und das Unmögliche. In *Publikationen der Deutschen Gesellschaft für Photogrammetrie und Fernerkundung*, vol. 16, 457–464. 21

MAYER, H. (2007b). Efficiency and Evaluation of Markerless 3D Reconstruction from Weakly Calibrated Long Wide-Baseline Image Loops. In *8th Conference on Optical 3-D Measurement Techniques*, vol. II, 213–219. 21

MAYER, H. & BARTELSEN, J. (2008). Automated 3D Reconstruction of Urban Areas from Networks of Wide-Baseline Image Sequences. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. (37) 5, 633–638. 80

MAYER, H., BARTELSEN, J., HIRSCHMÜLLER, H. & KUHN, A. (2012). Dense 3D reconstruction from Wide Baseline Image Sets. In F. Dellaert, J.M. Frahm, M. Pollefeys, L. Leal-Taixé & B. Rosenhahn, eds., *Theoretical Foundations of Computer Vision*, vol. 7474 of *Lecture Notes in Computer Science*, 285–304, Springer. 96

MICROSOFT (2010). Bing Maps 3D Control Has Been Discontinued. http://social.msdn.microsoft.com/Forums/en/vemapcontroldev/threads. 4

MICROSOFT (2011). Photosynth - Capture your World in 3D. http://photosynth.net/. 26

MICROSOFT (2012). Bing Maps Homepage. http://www.bing.com/maps/. 4

MIKOLAJCZYK, K. & SCHMID, C. (2004). Scale and Affine Invariant Interest Point Detectors. *International Journal of Computer Vision*, **60**, 63–86. 29, 32, 36, 37, 42, 43

MIKOLAJCZYK, K. & SCHMID, C. (2005). A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **27**, 1615–1630. 29

MIKOLAJCZYK, K., TUYTELAARS, T., SCHMID, C., ZISSERMAN, A., MATAS, J., SCHAFFALITZKY, F., KADIR, T. & VAN GOOL, L. (2005). A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, **65**, 43–72. 22, 23, 24, 29, 31, 33, 56, 75

MOREL, J. & YU, G. (2009). ASIFT: A New Framework for Fully Affine Invariant Image Comparison. *SIAM Journal on Imaging Sciences*, **2**, 438–469. 24, 32, 42, 43, 56, 89

NISTÉR, D. (2003). An Efficient Solution to the Five-Point Relative Pose Problem. In *Computer Vision and Pattern Recognition*, vol. II, 195–202. 2, 11, 12

NISTÉR, D. & STEWENIUS, H. (2006). Scalable Recognition with a Vocabulary Tree. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR'06, 2161–2168. 2

POLLEFEYS, M., VERGAUWEN, M. & VAN GOOL, L. (2000). Automatic 3D Modeling from Image Sequences. In *International Archives of Photogrammetry and Remote Sensing*, vol. (33) B5/2, 619–626. viii, 1, 7

POLLEFEYS, M., VAN GOOL, L., VERGAUWEN, M., VERBIEST, F., CORNELIS, K. & TOPS, J. (2004). Visual Modeling with a Hand-Held Camera. *International Journal of Computer Vision*, **59**, 207–232. viii, 1, 25

POLLEFEYS, M., NISTÉR, D., FRAHM, J.M., AKBARZADEH, A., MORDOHAI, P., CLIPP, B., ENGELS, C., GALLUP, D., KIM, S.J., MERRELL, P., SALMI, C., SINHA, S., TALTON, B., WANG, L., YANG, Q., STEWÉNIUS, H., YANG, R., WELCH, G. & TOWLES, H. (2008). Detailed Real-Time Urban 3D Reconstruction from Video. *International Journal of Computer Vision*, **78**, 143–167. viii, 1, 25

SCHAFFALITZKY, F. & ZISSERMAN, A. (2002). Multi-view Matching for Unordered Images Sets, or "How Do I Organize My Holiday Snaps?". In *Seventh European Conference on Computer Vision (ECCV'02)*, vol. I, 414–431. 25, 96

SCHÖDLBAUER, A. (1995). Rechenformeln und Rechenbeispiele zur Landesvermessung II. Herbert Wichmann Verlag Karlsruhe. 50

SNAVELY, N. (2010). Bundler: Structure from Motion for Unordered Image Collections. `http://phototour.cs.washington.edu/bundler/`. 25

STRECHA, C., VON HANSEN, W., VAN GOOL, L.J., FUA, P. & THOENNESSEN, U. (2008). On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery. In *21st IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. 10

STRECHA, C., PYLVANAINEN, T. & FUA, P. (2010). Dynamic and Scalable Large Scale Image Reconstruction. In *23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*. 27, 28, 36

TORR, P. (1998). Geometric Motion Segmentation and Model Selection. *Philosophical Transactions Royal Society of London A*, **356**, 1321–1340. 14

TORR, P. & ZISSERMAN, A. (1997). Robust Parametrization and Computation of the Trifocal Tensor. *Image and Vision Computing*, **15**, 591–605. 44

VISUAL GEOMETRY GROUP, O. (2003). Affine covariant regions datasets. http://www.robots.ox.ac.uk/~vgg/data/data-aff.html. viii, 22, 24, 25, 55

WINCK, B. (2008). 3D-Welten für die Bundeswehr. *Wehrwissenschaft Forschung und Technologie-Jahresbericht 2008*, 46–47. 4

WU, C. (2007). SiftGPU: A GPU Implementation of Scale Invariant Feature Transform (SIFT). http://cs.unc.edu/~ccwu/siftgpu. 19, 35, 37, 91, 94

WU, C., CLIPP, B., LI, X., FRAHM, J.M. & POLLEFEYS, M. (2008). 3D Model Matching with Viewpoint-Invariant Patches (VIP). In *21st IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. 57, 63

YAO, J. & CHAM, W.K. (2007). Robust Multi-View Feature Matching from Multiple Unordered Views. *Pattern Recognition*, **40**, 3081–3099. 96

ZADLO, S., WIEBROCK, I. & REINHARDT, W. (2010). 3D Modell der Universität der Bundeswehr München. http://www.unibw.de/inf4/professuren/geoinformatik/3d-modell-der-unibwm. 2