UNIVERSITÄT
KOBLENZ · LANDAU
Fachbereich 4: Informatik

Modellbasierte Posebestimmung aus 2-D/3-D SIFT-Korrespondenzen

Diplomarbeit
zur Erlangung des Grades
Diplom-Informatiker
im Studiengang Computervisualistik

vorgelegt von

Matthias Ebert

Betreuer: Dipl.-Inform. Peter Decker, Institut für Computervisualistik, Fachbereich Informatik, Universität Koblenz-Landau
Erstgutachter: Dipl.-Inform. Peter Decker, Institut für Computervisualistik, Fachbereich Informatik, Universität Koblenz-Landau
Zweitgutachter: Prof. Dr.-Ing. Dietrich Paulus, Institut für Computervisualistik, Fachbereich Informatik, Universität Koblenz-Landau

Koblenz, im Dezember 2009

Kurzfassung

Die Ermittlung der Position und Orientierung einer Kamera aus Punktkorrespondenzen zwischen 3D-Positionen und deren Bildpositionen ist im Rechnersehen unter dem Begriff *Poseschätzung* bekannt. Viele moderne Anwendungen profitieren von dem Wissen über die Lage einer Kamera im Raum zum Zeitpunkt der Bildentstehung. Für eine robuste Schätzung der Pose wird in dieser Arbeit zunächst anhand eines Stereoalgorithmus aus einer Bildserie ein Modell in Form einer Menge von SIFT-Merkmalen erstellt. Bei der Modellerstellung kommt eine handelsübliche monokulare Kamera zum Einsatz, die frei Hand geführt werden kann. Es ist dafür kein Wissen über die Position der Kamera während der Modellerstellung nötig. In einem zweiten Schritt wird die Pose einer Kamera bestimmt, deren Bild teilweise Inhalte des zuvor erstellten Modells aufweist. Die Zuordnungen der im Bild gefundenen SIFT-Merkmale zu den Modellmerkmalen mit bekannter 3D-Position bilden die Basis der linearen Optimierungsverfahren, die zur Lösung des Poseproblems angewandt werden. Das System beruht dabei auf einer zuvor kalibrierten Kamera und der manuellen Selektion geeigneter SIFT-Merkmale zur Initialisierung der Epipolargeometrie während des Modellaufbaus.

Abstract

The determination of a cameras position and orientation from point correspondences between 3d-positions and their image positions in computervision is known as *pose estimation*. Many modern applications benefit from the knowledge about the cameras absolute orientation in the reference frame at the time of image formation. To this extend a model is built from a sequence of images using structure-from-motion techniques and SIFT features. The model is built from a single off-the-shelf monocular camera which can be moved freehand. No a priori knowledge of the cameras position is needed while model construction. In a second step the pose of a camera wich shows partial content of the constructed model is computed. Mapping these model features and the features of the new image leads to the 3d-2d-correspondences which are the basis of linear optimization methods for solving the camera pose. The system relies on a precalibrated camera and a manual selection of adequate SIFT features for initial epipolar geometry estimation while model construction.

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Vereinbarung der Arbeitsgruppe für Studien- und Abschlussarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Mit der Einstellung dieser Arbeit in die Bibliothek bin ich einverstanden.	ja ☒	nein ☐
Der Veröffentlichung dieser Arbeit im Internet stimme ich zu.	ja ☒	nein ☐

Koblenz, den 17. Dezember 2009

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

In vielen Anwendungen ist das Wissen über die Position des Betrachters in seiner Umgebung von Nutzen. In *Augmented Reality* (AR) Szenarien wird es durch die Bestimmung der Betrachtungsposition erst überhaupt möglich, virtuell überlagerte Objekte korrekt zu positionieren und die vorherrschende Beleuchtungssituation darzustellen. Die meisten dieser Systeme basieren auf Markierungen (sogenannten *fiducials*) in der Welt. Sie stellen eine Referenz dar, die im Kamerabild verfolgt wird und zur Bestimmung der Position des Beobachters herangezogen wird.
In der Robotik ist die Lage des Systems in seiner Umgebung von hoher Bedeutung. Unter dem Begriff *SLAM* (simultaneous localization and mapping) wird das Problem der Selbstlokalisierung und gleichzeitiger Kartografierung der Umwelt des Roboters verstanden. Durch die Bestimmung der korrekten Lage im Raum und das Wissen über die Umgebung können kritische Situationen schon im Vorfeld erkannt und vermieden werden.

In dieser Arbeit wird ein Verfahren aufgezeigt, das mittels eines zuvor erstellten Modells und natürlichen Bildmerkmalen das Problem der eigenen Posebestimmung löst. Unter der Pose werden dabei die externen Kameraparameter, also ihre Rotation und Translation im Raum verstanden. Künstliche Markierungen in der Welt entfallen, da Korrespondenzen in dem Modell der rigiden Szene ermittelt werden. Der hier verwendete Ansatz lässt sich in zwei Schritte unterteilen: Ein Modell der Szene wird in einem ersten Schritt anhand eines Stereo-Algorithmus erstellt. Anschließend wird die gewonnene Modellinformation dazu genutzt, die Poseparameter einer neuen Ansicht der Szene zu bestimmen. Als Eingabemodalität dient eine handelsübliche monokulare Kamera, die freihändig in der Welt geführt wird. Moderne Digitalkameras sind kostengünstig, portabel, liefern einen hohen Detailgrad in ihren Aufnahmen und eignen sich deshalb für die Lösung des Poseproblems.

Die vorliegende Arbeit ist dabei in folgende Kapitel gegliedert:
In Kapitel 2 wird ein Überblick über die bestehenden Verfahren zur modellbasierten Posebestimmung gegeben. Kapitel 3 erläutert grundlegende Verfahren, die in der Arbeit zur Anwendung kommen. Es wird dabei auf die Extraktion der Bildmerkmale und das Problem der nächsten Nachbarschaftssuche solcher Merkmale eingegangen. Das darauf folgende Kapitel beschreibt den Ansatz, der hier zu der Lösung des Poseproblems gewählt wurde. Das Kapitel gliedert sich in die Schritte, die dazu nötig sind, das Modell der Szenerie zu erstellen und den daran anschließenden Verfahren der Posebestimmung.
Die praktische Vorgehensweise während der Erstellung des Modells und die Ergebnisse des Ansatzes sind in Kapitel 5 beschrieben. Kapitel 6 bildet eine Zusammenfassung und zeigt einige Möglichkeiten zur weiteren Optimierung des Verfahrens auf.
Im Anhang werden relevante mathematische Verfahren sowie ausgewählte Implementationsdetails dieser Arbeit erläutert.

Kapitel 2

Modellbasierte Posebestimmung

Die bestehenden Ansätze zur modellbasierten Posebestimmung unterscheiden sich sowohl in der Vorgehensweise der Modellerstellung als auch in den verwendeten Methoden zur Bestimmung der Kamerapose. Einen methodischen Überblick zur monokularen modellbasierten Poseschätzung verschafft [LF05]. Der Fokus der vorliegenden Arbeit richtet sich dabei auf die punktbasierten Verfahren.
Eine große Anzahl der punktbasierten Verfahren verwendet die Bildmerkmale des Kanade-Lucas-Tomasi Feature Tracker (KLT) [KL81] und des Harris Corner Detector [HS88]. Diese Bildmerkmale sind jedoch auf eine geringe Bewegung (*Baseline*) zwischen den Aufnahmen beschränkt. Speziell bei der Modellerstellung bedeutet das, dass eine dichte Anzahl an Aufnahmen der Szene notwendig ist. Die Grenzen der Poseschätzung einer neuen Aufnahme werden dadurch ebenfalls stark limitiert. Die natürlichen Punktmerkmale der Scale Invariant Feature Transform (SIFT) [Low04] weisen eine deutlich größere Invarianz gegenüber Blickwinkeländerungen auf und sind daher stabiler über eine weite Translation zwischen zwei Kameraaufnahmen wiederzuerkennen. Sie bilden die Basis für den hier verwendeten Ansatz zur modellbasierten Poseschätzung.
Ähnlich zu [BYY+08] wird ein sequentieller Ansatz der Modellerstellung gewählt, der von einer geordneten Bildserie ausgeht. Das erste Projektionsmatrizenpaar wird durch die Epipolargeometrie des ersten Bildpaares bestimmt. Weiterführende Bilder (*Frames*) werden direkt anhand der 3D-2D-Korrespondenzen in das Modell eingespeist.
Zur Bestimmung der Kamerapose wird ein Gleichungssystem aufgestellt, dessen Unbekannte die Translation und Rotation zwischen den Weltpunkten und den Bildpunkten darstellen. Jede 3D-2D-Korrespondenz fügt eine Gleichung in das Gleichungssystem ein. Es gibt zwei grundsätzliche Herangehensweisen, dieses zu lösen: lineare (direkte) Optimierungsverfahren und nichtlineare Optimierungsverfahren.
Lineare Verfahren zeichnen sich dabei durch ihre besondere Geschwindigkeitseffi-

zienz in der Berechnung aus, sind aber oft ungenauer als nichtlineare Verfahren. Ein normalverteiltes Bildrauschen wird durch lineare *Least-Squares-Methoden* gut ausgeglichen. Ausreißer verfälschen die Schätzung maßgeblich. Um deren Robustheit zu erhöhen, werden lineare Optimierungsverfahren meist mit dem RANdom SAmple Consensus (RANSAC) [FB81] kombiniert.
Ein lineares Verfahren bei Punktkorrespondenzen ist die *Direct Linear Transform* (DLT). Sie stammt ursprünglich aus der Fotogrammetrie und wurde von [Fau93] und [HZ03] in das Gebiet des Rechnersehens eingeführt. Die DLT schätzt die Kamerarotation und Translation in einer geschlossenen Form. Zur Schätzung der 12 Kameraparameter werden mindestens 6 Punktkorrespondenzen benötigt. Zur Steigerung der Robustheit schlägt [Har98] eine Normalisierung der Messwertmatrix vor.
Eine Methode, die mit nur 3 Punktkorrespondenzen auskommt wird in [FB81] als P3P (*Perspective 3 Point*) präsentiert. Bei bekannter Kalibrier-Matrix wird die Distanz von 3 Punkten aus Kamerasicht ermittelt. Ein Gleichungssystem wird aus den Abständen der Weltpunkte untereinander und dem eingeschlossenen Blickwinkel zwischen den Punkten aufgestellt. Dessen Lösung ist jedoch mit 3 Punktkorrespondenzen nicht eindeutig bestimmbar, es ergeben sich bis zu 4 mögliche Konfigurationen. [QL99] greift das Problem der Mehrdeutigkeit auf und stellt eine Methode dar, die mit 4 Punktkorrespondenzen (P4P) eine eindeutige Lösung gewährleistet.
[Fio01] stellt ein Verfahren vor, das mit 6 Punktkorrepsondenzen unter Berücksichtigung der Orthonormalität einer Rotation die Kameramatrix schätzt. Im Gegensatz zur DLT werden die Rotation und Translation isoliert voneinander betrachtet. Es wird zunächst die Tiefe der Weltpunkte aus Kamerasicht bestimmt, danach der gemeinsame Skalierungsfaktor des Modells errechnet und die optimale Rotation zwischen den Bildpunkten und Weltpunkten eingepasst. Abschließend wird die Translation der Kamera ermittelt.
Den zweiten Vertreter zur Poseschätzung bilden die nichtlineare Verfahren. Sie sind genauer, jedoch in der Berechnung wesentlich aufwändiger. Diese Vorgehensweisen benötigen eine initiale Schätzung (meist ein lineares Verfahren) und lösen ein nichtlineares Gleichungssystem iterativ. Die Initialisierung muss nahe an der optimalen Lösung liegen, ansonsten konvergieren nichtlineare Verfahren nicht schnell genug oder im schlechtesten Falle nie. Sie gelten ebenfalls als sehr störanfällig gegen Ausreißer in den Daten, die zur Optimierung herangezogen werden. Ein etablierter Vertreter ist die *Levenberg-Marquardt-Optimierung* [MNT04], die die Basis des *Bundle Adjustments* [TMHF00] ist. Bundle Adjustment (BA) optimiert in einem globalen Schritt sowohl die Struktur der Szene als auch die Bewegungsparameter sämtlicher Kameras und kommt in [VL04], [GL06], [BYY+08] zum Einsatz.

Kapitel 3

Grundlegende Verfahren

3.1 Extraktion der Bildmerkmale

In dieser Arbeit kommen die lokalen Bildmerkmale der Scale Invariant Feature Transform (SIFT) zum Einsatz. SIFT ist ein Algorithmus, der stabile Interessenspunkte auf einem Eingabebild detektiert und sie in einem hochgradig unterscheidbaren Deskriptorvektor repräsentiert. Das Verfahren wurde erstmals 1999 von David Lowe in [Low99] vorgestellt und kontinuierlich weiterentwickelt. In [Low04] wird dessen aktueller Entwicklungsstand beschrieben.
Die Scale Invariant Feature Transform erwartet als Eingabe ein monochromes Grauwertbild in 8-bit PGM-Format und liefert eine Menge von Merkmalen (auch Features genannt), die sich durch ihre Eigenschaften in vielen Bereichen der Bildverarbeitung bewährt haben.
Die herausragenden Eigenschaften der SIFT-Features und derer Deskriptoren hierbei sind:

Rotationsinvarianz: SIFT-Features werden in einer gegen Rotation invarianten Weise beschrieben. Sie sind in einer beliebig gedrehten Ansicht stabil wiederzuerkennen.

Skalierungsinvarianz: Sie sind invariant gegen eine Skalierung, was einen robusten Vergleich von Bildpunkten gewährleistet, die einem Objekt entspringen, das in zwei Aufnahmen unterschiedlich groß erscheint.

Subpixelgenaue Lokalisierung: Die Bildposition der Interessenspunkte ist mit subpixelgenauer Präzision beschrieben.

Anzahl: SIFT produziert in einem durchschnittlich texturiertem Bild stets eine hohe Anzahl von Merkmalen.

Robustheit: Die Merkmale gelten dabei als sehr robust und unanfällig gegen Bildrauschen.

Unterscheidbarkeit: Sie sind darüber hinaus in einem hochgradig unterscheidbarem Deskriptorvektor beschrieben, wodurch sie besonders geeignet sind, das Korrespondenzproblem des Rechnersehens stabil zu lösen.

Invarianz gegen affine Transformation: Sie sind bis zu einem gewissen Grad invariant gegen affine Transformationen.

Invarianz gegen Beleuchtungsänderung: Ebenfalls eine partielle Invarianz gegen Beleuchtungsänderung wird durch die Eigenschaften des Deskriptorvektors erreicht.

Im Folgenden sind die wichtigsten Stationen der Scale Invariant Feature Transform, die zu den eben beschriebenen Eigenschaften führen, zusammengefasst.

3.1.1 Erstellung des Skalenraums

Der erste Schritt der Scale Invariant Feature Transform besteht in der Erzeugung des Skalenraums. Das Konzept des Skalenraums und dessen Bedeutung für das Rechnersehen wurde in [Lin94] abgehandelt. Ein Eingabebild $I(x, y)$ wird zu Beginn mit einem Gaußkern variabler Glättungsfaktoren σ, $G(x, y, \sigma)$, wiederholt gefaltet und der Skalenraum als eine kontinuierliche Funktion $L(x, y, \sigma)$ definiert:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{3.1}$$

Durch die Differenzbildung zweier solch aufeinanderfolgender Skalen, die durch einen konstanten Multiplikationsfaktor k getrennt sind, ergibt sich eine Annäherung der normalisierten Laplaceebenen [Low04], die die stabilsten Merkmalskandidaten liefert:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{3.2}$$

Eine solche Differenzbildung wird als Difference-Of-Gaussian (DOG) bezeichnet. Sind fünf Gaußebenen erstellt, werden sie in einer *Oktave* zusammengefasst, um die Hälfte in der Auflösung skaliert und wiederum in einer neuen Oktave gruppiert. Dies führt zu 4 DOG-Bildern pro Oktave.
Das Eingabebild wird also mit steigender Skala immer stärker geglättet und mit zunehmender Oktave in der Auflösung reduziert. Details gehen verloren, wobei dominante Strukturen im Bild enthalten bleiben. Diese Repräsentation entspricht

dem allgemeinen Wahrnehmungsempfinden von Objekten, die mit größerer Entfernung des Betrachters (respektive mit kleiner werdender Skalierung) in ihrem Detailgrad abnehmen, in ihrer äußeren Form jedoch erhalten bleiben.
Durch die Skalenraumdarstellung eines Bildes wird eine Skalierungsinvarianz der lokalen Bildmerkmale erreicht, da Merkmalspunkte auf allen Skalen eines Eingabebildes gesucht werden.

Abbildung 3.1 veranschaulicht die Konstruktion des Skalenraums durch DOG-Bildung der mit zunehmenden σ geglätteten Bildebenen.

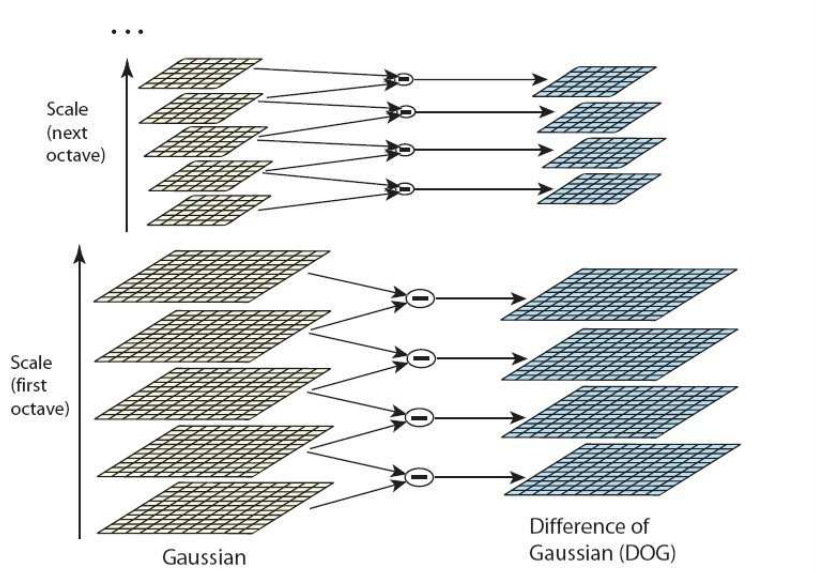


Abbildung 3.1: Skalenraum, Quelle: [Low04]

3.1.2 Detektion lokaler Extremstellen im Skalenraum

Ist der Skalenraum erstellt, werden mögliche Merkmalskandidaten als Extremstellen im Skalenraum gesucht. Dazu werden die Werte der 8 Nachbarn jedes Punktes seiner Skala und die 9 Nachbarn der jeweils angrenzenden Skalen verglichen. Ist ein Punkt das Maximum oder Minimum dieser 26 Werte, wird er als möglicher Merkmalskandidat betrachtet und geht in weitere Berechnungen mit ein. Abbildung 3.2 veranschaulicht die inspizierte Nachbarschaft der angrenzenden Skalen.

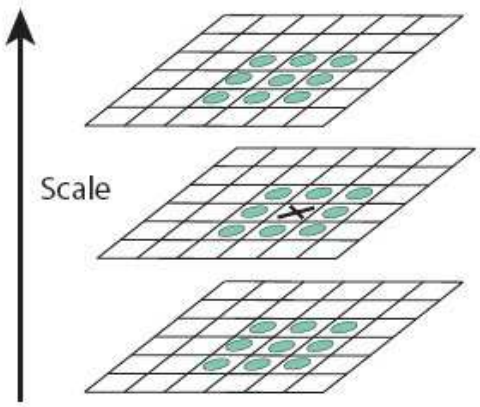


Abbildung 3.2: Extremstellensuche im Skalenraum, Quelle: [Low04]

3.1.3 Subpixelgenaue Lokalisierung der Merkmalspunkte

In einem nächsten Schritt werden mögliche Kandidaten, die aufgrund von niedrigem Kontrast anfällig gegen Bildrauschen sind eliminiert. Da der DOG-Operator stark auf Kanten reagiert, werden viele Extremstellen im Skalenraum auf Kanten liegen. Diese Punkte sind schlecht lokalisierbar und werden deshalb ebenfalls aussortiert. Dabei werden die Eigenwerte der Hessematrix der Bildregion inspiziert. Falls eine große Differenz zwischen ihnen besteht, wird der Punkt auf einer Kante angenommen und verworfen. Dies trägt wesentlich zur Stabilität der verbleibenden Merkmale bei.
Da besonders auf hohen Skalen eine genaue Lokalisierung des Merkmals durch die sehr geringe Auflösung ungenau und schwierig ist, wird eine dreidimensionale quadratische Funktion in die Umgebung der Merkmalsposition eingepasst und die subpixelgenaue Position des Merkmals aus deren Eigenschaften ermittelt.

3.1.4 Zuweisung der Orientierung

Ist ein Merkmal gefunden, wird die dominante Orientierung der Nachbarschaft bestimmt. Ziel ist es, eine Rotationsinvarianz des Merkmals zu erreichen, indem alle folgenden Informationen über die lokale Nachbarschaft an dieser dominanten Orientierung normalisiert sind. Dazu wird die Gradientengröße

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \quad (3.3)$$

und Richtung

$$\Theta(x,y) = \tan^{-1}((L(x+1,y) - L(x-1,y))/(L(x,y+1) - L(x,y-1)) \quad (3.4)$$

der lokalen Nachbarschaft eines Merkmals berechnet, wobei $\Theta \in [-\pi, \pi]$. Die Gradientengrößen werden in ein Orientierungshistogramm, das aus 36 Urnen besteht, jede Urne einen Bereich von 10° bedeckend, sortiert. Die Orientierung liegt in Bogenmaß vor und muss deshalb für eine korrekte Zuordnung in Winkelmaß umgerechnet werden. Eine dominante Orientierung für einen Merkmalspunkt ist nun an dem Maximum der Orientierungshistogramme gefunden. Diesem Merkmal wird damit die dominante Orientierung Θ des jeweiligen Orientierungshistogramms zugeordnet [1]. Liegen weitere Maxima innerhalb von 80% des Scheitelpunktes in einem der Orientierungshistogramme vor, so werden zusätzliche Merkmale mit eben derselben lokalen Bildposition und einer neuen Hauptorientierung Θ erzeugt.

3.1.5 Erstellung des Deskriptors

Bisher wurden lediglich die Position und Orientierung eines Merkmals bestimmt, aber keine Aussage über den Bildinhalt der lokalen Nachbarschaft getroffen. Um Merkmale vergleichen zu können, müssen sie möglichst eindeutig beschrieben werden. Bei der Scale Invariant Feature Transform wird deshalb zu jedem Punkt dessen Nachbarschaft untersucht und deren Bildinhalte in einem Deskriptorvektor repräsentiert. Die besten Ergebnisse wurden dabei mit einem 128-dimensionalen Deskriptor erzielt [Low04].
Bei der Erstellung des Deskriptorvektors werden sämtliche Histogramme und Bildkoordinaten an der eben erschlossenen Hauptorientierung ausgerichtet und sind dadurch in einer rotationsinvarianten Art beschrieben.
In einer Nachbarschaft von 16×16 Pixel um die Position des Merkmals werden die Gradientengrößen (siehe 3.3) der jeweiligen Pixel errechnet und mit einem Gaußfenster gewichtet, sodass die Pixelpositionen mit zunehmender Entfernung einen abnehmenden Beitrag in der Gradientengröße entsprechen. In Abbildung 3.3 ist die Gewichtung durch das kreisförmige Gaußfenster auf der linken Seite illustriert. Die Absicht der Gewichtung ist es, dass kleine Änderungen in der Position des Merkmals auch nur zu kleinen Änderungen in der Repräsentation des Deskriptors

[1] Es wird eine parabolische Funktion in das Maximum und dessen 3 nächsten Werten eingepasst, um die Diskretisierung in 10° Schritte zu interpolieren.

führen.
Neben den Gradientengrößen werden analog zu 3.4 deren Orientierungen errechnet. Für die 4 Quadranten um den lokalen Merkmalspunkt werden je 4 Orientierungshistogramme mit nun 8 Urnen erstellt. Jede der 8 Urnen deckt einen Bereich von 45° ab. Die Information aus den Gradientengrößen der 16er Nachbarschaft wird in das passende Orientierungshistogramm (je nach Quadrant, Lage und Orientierung des Gradienten) akkumuliert. Abbildung 3.3 zeigt dies veranschaulicht an einer 8×8 Nachbarschaft mit entsprechend 4 Histogrammen zu jeweils 8 Urnen.
Die letztendliche Dimension des SIFT-Deskriptors ergibt sich aus den 4 Quadranten mit je 4 Orientierungshistogrammen, jedes durch 8 Urnen charakterisiert, also insgesamt $4 * 4 * 8 = 128$ Dimensionen. Die Histogramme werden in dem Deskriptorvektor kombiniert und bilden dessen Inhalt.

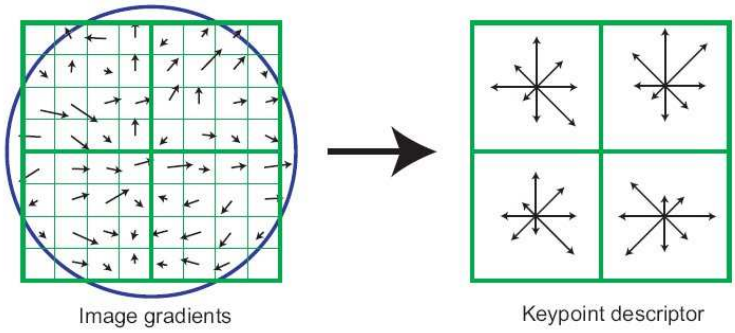


Abbildung 3.3: Gradienten in der Umgebung eines Merkmals (links), Orientierungshistogramme (rechts), Quelle: [Low04]

Um eine partielle Invarianz gegenüber Effekten, die durch Beleuchtungsänderungen entstehen, herzustellen, wird der Deskriptor in einem letzten Schritt normalisiert. Durch die Normalisierung werden lineare Beleuchtungseffekte, wie Kontrast- oder Helligkeitsschwankungen beseitigt. Um mit nichtlinearen Beleuchtungsänderungen, wie z.B. einer Sättigung der Kamera, umzugehen, werden alle Elemente des Deskriptors, die größer als 0,2 sind auf 0,2 gesetzt und in einem letzten Schritt der Deskriptorvektor abermals normalisiert.
Für jedes Eingabebild werden die zuvor beschriebenen Schritte durchgeführt und das Ergebnis in einer Merkmalsliste abgespeichert. Ein typischer Aufbau einer solchen Liste ist in Abbildung 3.4 dargestellt. Die Aufzählung beginnt mit der insgesamten Anzahl an detektierten Schlüsselpunkten in einem Bild, gefolgt von der Dimension der Deskriptoren (im Allgemeinen 128). Es wird für jeden Merkmalspunkt dessen subpixelgenaue Position in Y-Pixelkoordinaten und X-Pixelkoordinaten so-

wie die Skala und Orientierung notiert. Die charakteristischen Einträge des zugehörigen Deskriptors folgen.

```
<Anzahl der Merkmale> <Dimension der Deskriptoren>

<Y-Pixelposition 1> <X-Pixelposition 1> <Skala 1> <Orientierung 1>
    <Deskriptor 1>

<Y-Pixelposition 2> <X-Pixelposition 2> <Skala 2> <Orientierung 2>
    <Deskriptor 2>

...

<Y-Pixelposition N> <X-Pixelposition N> <Skala N> <Orientierung N>
    <Deskriptor N>
```

Abbildung 3.4: Aufbau einer Merkmalsliste

3.2 Nachbarschaftssuche zwischen Merkmalen

Auf der Basis der Deskriptorvektoren lassen sich die Merkmale nun gegenseitig zuordnen. Dies ist ein erster Schritt zur Lösung des Korrespondenzproblems. Um den nächsten Nachbarn eines Punktes zu bestimmen, muss die Menge der restlichen Interessenspunkte durchlaufen und ein Maß für den Abstand zweier Punkte errechnet werden.
Die Menge an extrahierten Bildmerkmalen ist durch die Dimension ihrer Deskriptorvektoren Teil eines 128-dimensionalen Vektorraums $X = \mathbb{Z}^{128}$. Bei der gesamten Menge an Punkten $I = \{i_1, ..., i_n\}$ des Vektorraums X ist nun der Punkt i gesucht, der einem gegebenem $j \in X$ der nächste ist. Seien d_i und d_j die Deskriptoren der jeweiligen Punkte, so ist der nächste Nachbar (und somit eine mögliche Korrespondenz) eines Abfragepunktes definiert als derjenige Punkt, dessen Deskriptor den geringsten Abstand zu dem Deskriptor des Abfragepunktes aufweist[2].
Es ist zunächst die Wahl der Distanzfunktion zu bestimmen. Eine Distanzfunktion $distance(i, j) : I \times I \rightarrow \mathbb{R}$ der Punkte $i \in I$ und $j \in I$ ist definiert, falls gilt:

[2]Im Folgenden wird ein Punkt i äquivalent zu seinem Deskriptor d_i angesehen.

$$distance(i,j) = distance(j,i) \tag{3.5}$$
$$distance(i,j) \geq 0 \tag{3.6}$$
$$distance(i,i) = 0 \ \forall i,j \in I \tag{3.7}$$

Die in dieser Arbeit verwendete Distanzfunktion ist der euklidische Abstand (l_2-Norm) zweier Punkte i und j in einem 128-dimensionalen Vektorraum:

$$distance(i,j) = \sqrt{\sum_{k=1}^{128}(i_k - j_k)^2} \tag{3.8}$$

Auf Basis der Distanzfunktion wird zur Bestimmung des nächsten Nachbarn folglich der Datensatz der Punkte durchlaufen. Einige Ansätze hierzu werden in 3.2.1 und 3.2.2 aufgezeigt.

3.2.1 Exakte lineare Suche

Der klassische Ansatz der nächsten Nachbarschaftssuche ist die lineare Suche. Dabei wird ein gegebener Abfragepunkt mit jedem Punkt der Datenmenge verglichen, deren Distanz berechnet und der Punkt als nächster Nachbar angesehen, falls seine Distanz kleiner als das bisherige Minimum der Distanzen ist. Nach einem vollständigen Durchlauf der Datenmenge ist der tatsächliche nächste Nachbar zu einem Punkt gefunden.
Die lineare Suche hat pro Suchanfrage einen Aufwand von

$$\mathcal{O}(Nd) \tag{3.9}$$

bei einer Anzahl von N Punkten und der Dimension d des Vektorraums.
Es ist daher klar, dass bei steigender Anzahl an Merkmalen und der ohnehin hohen Dimension derer Deskriptoren eine lineare Suche schnell unpraktikabel wird.

3.2.2 Angenäherte Suche

Wegen der hohen Laufzeit der linearen Suche wurden Suchalgorithmen entwickelt, die den Vektorraum in Regionen unterteilen, in denen potentielle nächste Nachbarn zusammengefasst sind. Die Regionen werden in einer Datenstruktur indexiert, die während der Suchanfrage durchlaufen wird. Dabei wird in Kauf genommen, dass das Ergebnis der Suchanfrage nicht der tatsächliche nächste Nachbar eines Punktes ist, sondern diesem nur mit hoher Wahrscheinlichkeit entspricht. Verfahren dieser Art werden als *angenäherte Nachbarschaftssuche* bezeichnet. Bei der angenäherten

Nachbarschaftssuche ist ein Punkt $i \in X$ ein ε-nächster Nachbar eines gegebenen Punktes $j \in X$, wenn gilt: $dist(i,j) \leq (1+\varepsilon)dist(i^*,j)$ und i^* ist der tatsächliche nächste Nachbar. Dieses ε ist nur zu bestimmen, falls der tatsächliche NN (nächster Nachbar) bekannt ist. Es wird allerdings als gering angenommen, wenn der Suchlauf eine Mindestanzahl möglicher nächster Nachbarn verglichen hat.
Hochdimensionale Suchräume gelten wegen des *Fluchs der Dimensionen* als schwierig zu strukturieren. Der Fluch der Dimensionen bezeichnet die Tatsache, dass die Abstände zufällig erzeugter oder normalverteilter Daten mit zunehmender Dimension dazu tendieren immer äquidistanter zu werden. Punktmengen, die realen Bilddaten entspringen, weisen diese Tendenz jedoch nicht auf und sind daher gut zur lokalen Gruppierung geeignet.
Die gängigsten Verfahren nutzen als Indexierungsstruktur entweder Hashing-Funktionen (zum Beispiel *LSH* [GIM99]) oder binäre Suchbäume.
In der vorliegenden Arbeit wird die *Fast Library for Approximate Nearest Neighbor* (*FLANN*) verwendet [ML09]. FLANN inspiziert anhand einer Stichprobe die Beschaffenheit des Datenraumes und wählt automatisch zwischen zwei zur Verfügung stehenden Indexierungsmethoden die geeignete aus. Die Methoden sind:

- Multiple zufällige Kd-Trees: Statt wie bei Kd-Trees üblich, strikt an der mittleren Dimension zu teilen, werden die Bäume an einer der fünf Dimensionen mit der größten Varianz aufgespaltet. Diese Dimension wird jeweils zufällig gewählt. Es werden parallel mehrere Suchbäume erstellt.

- Hierarchischer K-Means-Tree: Der Datensatz wird um dessen Schwerpunkt geclustert und nach einer zuvor gewählten maximalen Anzahl der Kandidaten eines Clusters rekursiv verfeinert.

Beide Datenstrukturen werden anhand einer Priority-Queue, die die Reihenfolge der noch nicht besuchten Blätter nach deren Distanz hält, durchlaufen. Die gewünschte Präzision der Suchanfrage ergibt sich durch die Anzahl der untersuchten Punkte, die gleichzeitig das Abbruchkriterium der Anfrage darstellt.
Im Vergleich zu einer linearen Suchanfrage hat sich der Geschwindigkeitsvorteil als enorm erwiesen. Ein Faktor von 10^3 bei einer Präzision von über 80% wird in [ML09] angegeben.

Kapitel 4

Modellbasierte Pose aus SIFT-Korrespondenzen

In den folgenden Kapiteln wird der hier verwendete Ansatz der *Modellbasierten Posebestimmung aus 2-D/3-D SIFT-Korrespondenzen* erläutert. Es wird ein Modell einer rigiden Szene erstellt und die Pose einer Kamera durch robuste Zuordnung zwischen Modellmerkmalen und Merkmalen einer neuen Ansicht bestimmt. Das Problem der Poseschätzung einer Kamera ist in der Ermittlung ihrer externen Parameter in Bezug auf ein Referenzkoordinatensystem formuliert. Abbildung 4.1 verdeutlicht eine solche Transformation.

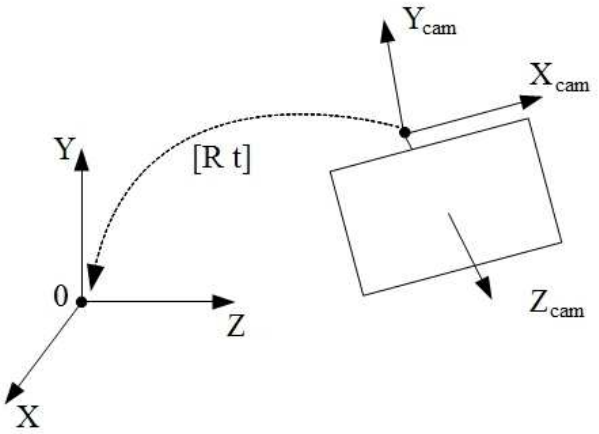


Abbildung 4.1: Transformation einer Kamera

Eine vollständige projektive Abbildung eines Weltpunktes $\tilde{\boldsymbol{X}}^{\mathrm{w}} = \begin{bmatrix} X & Y & Z & 1 \end{bmatrix}$ in einen Bildpunkt $\tilde{\boldsymbol{x}}^{\mathrm{i}} = \begin{bmatrix} x & y & 1 \end{bmatrix}$ setzt sich zusammen aus:

$$\tilde{\boldsymbol{x}}^{\mathrm{i}} \sim \boldsymbol{K}\boldsymbol{P}\tilde{\boldsymbol{X}}^{\mathrm{w}} \tag{4.1}$$

Die Kalibriermatrix $\boldsymbol{K}$ repräsentiert die intrinsischen Kameraparameter (siehe 4.1.1), $\boldsymbol{P} = \begin{bmatrix}\boldsymbol{R} & \boldsymbol{t}\end{bmatrix}$ bezeichnet die Rotation $\boldsymbol{R}$ und Translation $\boldsymbol{t}$ der Kamera in Weltkoordinaten und $\sim$ ist dabei die Gleichheit bis auf einen unbekannten skalaren Faktor. Die 3×4 Matrix $\boldsymbol{P}$ wird im Folgenden auch Projektionsmatrix oder Kameramatrix genannt. Sie beschreibt die *externe Orientierung* der Kamera im Raum und ist deshalb auch als *Kamerapose* bekannt.

Ziel dieser Arbeit ist es, die Poseparameter anhand von 2D-3D-Korrespondenzen zu bestimmen, was dem Wissen über Lage und Orientierung der Kamera zum Zeitpunkt der Bildentstehung entspricht.

Die einzelnen Schritte, die dazu nötig sind, werden in den nun folgenden Kapiteln ausgeführt.

4.1 Erstellung des Modells

In einem ersten Schritt wird anhand von SIFT-Korrespondenzen und einem Stereo-Algorithmus ein dreidimensionales Modell einer Bildserie in Form einer Punktewolke konstruiert. Das Ziel der Modellerstellung ist, mittels Struktur-Aus-Bewegungs-Algorithmen eine möglichst präzise Repräsentation der betrachteten Szene zu generieren und alle notwendigen Informationen für eine Poseschätzung bereitzustellen. Als Eingabedaten erwartet der Algorithmus eine Serie von Bildern und die Kalibrierparameter der verwendeten Kamera (siehe 4.1.1). Die Bilder können mit einer beliebigen Kamera erstellt werden, es ist jedoch eine gewisse Anzahl an extrahierten Merkmalen nötig, um Anschlussfehler zwischen zwei fortlaufenden Aufnahmen zu vermeiden[1]. Um Korrespondenzen zwischen Bildern zu finden, ist eine grundlegende Voraussetzung, dass sich teilweise Bildinhalte in aufeinanderfolgenden Frames wiederholen. Eine maximale Blickwinkeländerung ist durch die Struktur der Szene gegeben. Werte von bis zu circa 30° sind jedoch akzeptabel.
Mindestens zwei adjazente Bildaufnahmen sind nötig, um das Modell zu erstellen. Bei steigender Anzahl wird das Modell visuell dichter, enthält mehr Informationen und erlaubt einen größeren Betrachtungsraum für die anschließende Poseschätzung. Jede neue Ansicht eines bereits betrachteten Punktes bringt die zusätzliche Information aus dem aktuellen Blickwinkel mit ein (siehe 4.1.9).
Die in dieser Arbeit erstellten dreidimensionalen Modelle bestehen aus bis zu 43

[1] Die Anzahl der SIFT-Features hängt sowohl von der verwendeten Auflösung als auch von dem Bildinhalt ab.

Einzelbildern, die bis zu 360° Ansichten darstellen (siehe Kapitel 5). Die Szene wird als rigide und unbeweglich angesehen, mögliche dynamische Strukturen[2] werden als Ausreißer behandelt und verworfen.

Zusammenfassend lässt sich die Modellerstellung in die folgenden Schritte unterteilen, die für jedes Bildpaar sequentiell auszuführen sind:

1. Lokale invariante Merkmale werden von den Eingabebildern extrahiert.
2. Automatische Zuordnungen werden über fortlaufende Bildpaare bestimmt.
3. Korrekte Korrespondenzen werden anhand einer durch manuelle Korrespondenzen geschätzten Fundamental-Matrix ermittelt.
4. Die Kameramatrix des neuen Bildes wird errechnet.
5. Die korrekten Zuordnungen werden trianguliert und die Beziehung der Weltpunkte zu ihren Bildpunkten erfasst.

Abbildung 4.2 zeigt einen exemplarischen Durchlauf für eine Bildserie, dessen einzelne Stationen in den nächsten Kapiteln näher erläutert werden.

4.1.1 Kamerakalibrierung

In einem Kalibrierschritt nach der Methode von [Zha00] wurde im Vorfeld der Anwendung die $\boldsymbol{K}$-Matrix der jeweiligen Kamera, die deren intrinsischen Kameraparameter enthält, ermittelt. Mit Hilfe dieser Parameter wird die Extraktion der Essential-Matrix aus der Fundamental-Matrix (siehe 4.1.6) sowie die Umrechnung von Pixelkoordinaten in Bildkoordinaten vorgenommen. Desweiteren wurden bei der Kalibrierung der Kamera deren radiale Verzerrungsparameter ermittelt. Sie sind Basis für die radiale Entzerrung der Bilder, ein wichtiger Schritt, der auf sämtliche Eingabebilder a priori ausgeführt wird.
Die intrinsischen Kameraparameter einer Kalibrier-Matrix lassen sich beschreiben durch:

$$\boldsymbol{K} = \begin{bmatrix} \alpha & \lambda & \boldsymbol{u_0} \\ 0 & \beta & \boldsymbol{v_0} \\ 0 & 0 & 1 \end{bmatrix} \tag{4.2}$$

mit $(\boldsymbol{u_0}, \boldsymbol{v_0})$ den Koordinaten des Hauptpunktes der Kamera, α und β den Skalierungsfaktoren in Bildkoordinatenachsen $\boldsymbol{u}$ und $\boldsymbol{v}$ sowie einem Parameter λ, der die Schrägstellung des Bildsensors in Bezug auf das Kamerakoordinatensystem

[2] Zum Beispiel Personen, die durch das Bild laufen oder andere bewegliche Objekte.

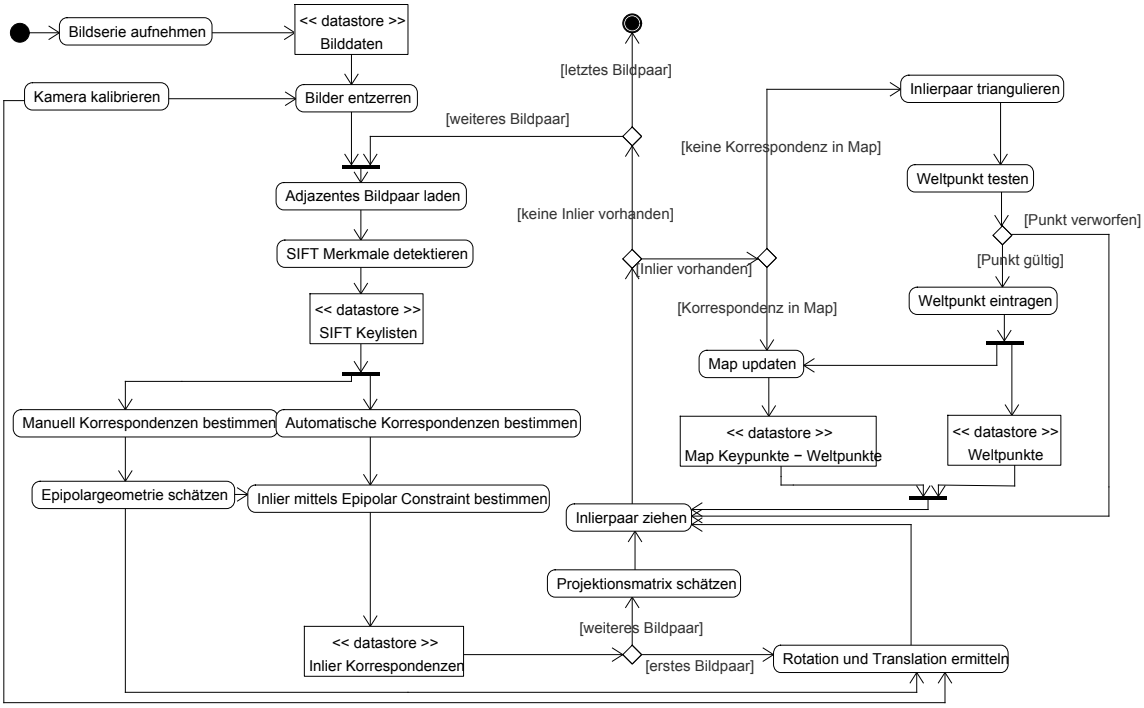


Abbildung 4.2: Aktivitätsdiagramm der Modellerstellung

darstellt.
Da diese Informationen an einigen Stellen während des Ablaufs der Modellerstellung und der Poseschätzung von Nöten sind, werden sie in Dateien gehalten und stehen somit jederzeit zur Verfügung.

4.1.2 Merkmalsdetektion

In der vorliegenden Arbeit werden die in 3.1 beschriebenen lokalen Bildmerkmale der Scale-Invariant-Feature-Transform verwendet. Sie zeichnen sich durch ihre hohe Unterscheidbarkeit und ihre Stabilität gegenüber Bildrauschen aus und sind daher eine gute Wahl, die spezifischen Probleme, die an diese Arbeit gestellt werden zu lösen.
Darüber hinaus produziert der SIFT-Algorithmus eine dichte Anzahl an Punkten in einem durchschnittlich texturiertem Bild und ist daher besonders geeignet für lineare Optimierungsverfahren, die von einer großen Eingabedatenmenge profitieren. Die hohe Anzahl an SIFT-Merkmalen eines Bildes erlaubt es außerdem, mit ausreißerbehafteten Daten eine robuste Schätzung durchzuführen, ohne die nötige Anzahl der Modellparameter zu unterschreiten.

Das Hauptkriterium für ihre Wahl sind jedoch die oben beschriebenen Eigenschaften der Invarianz gegenüber Rotation und Skalierung. Desweiteren sind sie in der Lage, unter Blickwinkelneigungen von bis zu 50° und einem zusätzlichen Bildrauschen von 4% verlässliche Zuordnungen zu gewährleisten [Low04].

4.1.3 Automatische Korrespondenzbestimmung der detektierten Merkmale

Sind die Merkmale durch ihren Deskriptorvektor beschrieben, gilt es auf der Basis dieser Information Korrespondenzen von Merkmalen zu finden, die in einer fortlaufenden Bildsequenz ein und demselben physischen Weltpunkt entspringen.
Ein 2D Bildpunkt in Pixelkoordinaten $\tilde{\boldsymbol{x}}_{ij}^{\mathrm{p}}$ entspricht der Projektion eines 3D Weltpunktes $\tilde{\boldsymbol{X}}_{j}^{\mathrm{w}}$ in ein Bild i :

$$\tilde{\boldsymbol{x}}_{ij}^{\mathrm{p}} \sim \boldsymbol{P}_i \tilde{\boldsymbol{X}}_{j}^{\mathrm{w}} \tag{4.3}$$

Die Projektionsmatrix bildet einen Weltpunkt auf einen Bildpunkt der jeweiligen Kamera ab und somit sind zwei Korrespondenzen durch ihre Projektionsmatrizen und deren Weltpunkt bestimmt.
Da aber genau diese Projektionsmatrizen nicht bekannt sind, muss die Korrespondenzsuche im Bildraum anhand der von SIFT annotierten Deskriptorvektoren erfolgen. Eine mögliche Korrespondenz zu einem gegebenem Bildpunkt ist dessen nächster Nachbar aus einer Menge an extrahierten Bildmerkmalen eines konsekutiven Frames (siehe 3.2).

Allein auf Basis der nächsten Nachbarschaftssuche eine Korrespondenzbestimmung als gültig anzusehen, wäre nicht ausreichend. Das Ergebnis der nächsten Nachbarschaftssuche ist nicht immer korrekt (siehe 3.2.2) und die tatsächlichen nächsten Nachbarn erweisen sich unter bestimmten Bedingungen als nicht stabil genug für eine weitere Verwendung. Deshalb wird das Ergebnis der nächsten Nachbarschaftssuche bisher als *mögliche* Korrespondenz angesehen, die weiteren Überprüfungen unterzogen werden muss, bevor sie als korrekt bezeichnet wird. Ein wichtiger Schritt zur Steigerung der Robustheit dieser Zuordnung ist folgend erklärt.

Die Stabilität eines Merkmals in Bezug auf Korrespondenzbestimmung hängt stark von der Beschaffenheit des Merkmalsraumes ab, in dem es sich befindet. Viele Punkte, die von Strukturen entstehen, die sich sehr selbstähnlich sind (wie zum Beispiel die Blätter eines Baumes, eine Fensterfront mit sich wiederholenden Elementen oder oft wiederkehrende Muster) haben viele mögliche Korrespondenzen, die einen gegenseitig sehr geringen euklidischen Abstand aufweisen. Selbst

der Punkt mit dem kleinsten euklidischen Abstand ist daher als nicht stabil anzusehen, da in einem nächsten Suchlauf ein sehr ähnlicher Punkt zurückgegeben werden könnte (im Falle der angenäherten Nachbarschaftssuche), der an einer anderen Bildposition lokalisiert ist. Solche falsch positiv angesehenen Zuordnungen werden als Ausreißer (auch *Outlier*) bezeichnet und sind unbedingt abzufangen. Als Synonym für eine korrekte Zuordnung wird in Zukunft auch der Begriff *Inlier* verwendet.
Um ungeeignete mögliche Korrespondenzen schon im Vorfeld zu eliminieren, werden stets die zwei nächsten Nachbarn (j_1 und j_2) eines Abfragepunktes (i) inspiziert. Ist das Verhältnis derer Abstände zu dem Referenzpunkt zu gering, deutet dies auf ein instabiles und selbstähnliches Merkmal hin. Der nächste Nachbar wird also nicht als Korrespondenz angesehen, sondern verworfen.
Das Verhältnis der Abstände wird als *distanceRatio* bezeichnet:

$$distanceRatio = \frac{distance(i, j_1)}{distance(i, j_2)} \tag{4.4}$$

Ist j_1 der nächste und j_2 der zweitnächste Punkt, so ist der Schwellwert r der *distanceRatio*: $0 < r < 1$. Je näher an 1, desto selbstähnlicher ist das Merkmal zu anderen Merkmalen, je näher an 0, desto unterscheidbarer ist es.
[Low04] gibt an, dass durch Aussortieren von Punkten deren Verhältnis $r > 0,8$ beträgt etwa 90% der falschen Zuordnungen eliminiert werden, wobei weniger als 5% der korrekten Zuordnungen fälschlicherweise verworfen werden. In dieser Arbeit werden Schwellwerte von 0,6 - 0,8 verwendet, je nach Bildinhalt sind diese anzupassen.

4.1.4 Manuelle Korrespondenzauswahl zur Schätzung der Epipolargeometrie

Das Korrespondenzproblem des Rechnersehens allein durch eine automatische Zuordnung zu lösen, ist wegen der noch immer vorhandenen Ausreißer nicht möglich. Ein geeignetes Mittel, deren Korrektheit zu steigern, besteht in einer Überprüfung der Daten durch geometrische Einschränkungen (siehe 4.1.5). Eine solche Überprüfung setzt jedoch das Wissen über die Epipolargeometrie zwischen den zwei Perspektiven voraus, aus denen die Merkmale detektiert wurden. Deshalb gilt es zunächst die Epipolargeometrie zwischen den Ansichten zu schätzen. Sie ist jedoch nur bei einer optimalen Beschaffenheit der Datenpunkte ohne weiteres zu bestimmen.
Daten, die nicht ausreichend sind, um die Epipolargeometrie mittels Struktur-Aus-Bewegung-Algorithmen eindeutig zu schätzen, werden als *degenerierte Daten*

bezeichnet. [FTZ99] kategorisiert das Problem der degenerierten Daten in zwei Bereiche. Bei Daten, die aus einer Bildfolge zweier Frames stammen, die sich in ihrer Bewegung als reine Rotation oder reine Translation unterscheiden, handelt es sich um eine Bewegungsdegeneration (*motion degeneracy*). Befinden sich die betrachteten Korrespondenzen zwischen zwei Kamerabildern auf einer planaren Ebene in der Welt oder sind auf eine kleine Region im Bild konzentriert, liegt eine Strukturdegeneration vor (*structure degeneracy*). In beiden Fällen ist das Bewegungsmodell lediglich durch die Homographie der Bilder bestimmt und die Eingabedaten reichen nicht aus, um die Tiefeninformation der Bildpunkte und die Position der Kameras zu bestimmen.
Abbildung 4.3 veranschaulicht beide Fälle anhand einer Aufnahmereihe eines Gebäudes. Zwischen den Kameras drei und vier liegt eine Bewegungsdegeneration vor. Die Bewegung zwischen den Kameras ist auf eine Rotation um das optische Zentrum reduziert und es liegt keine Translation vor. Die Situation zwischen den Kameras sieben bis neun skizziert eine Strukturdegeneration aufgrund der planaren Gebäudefront.

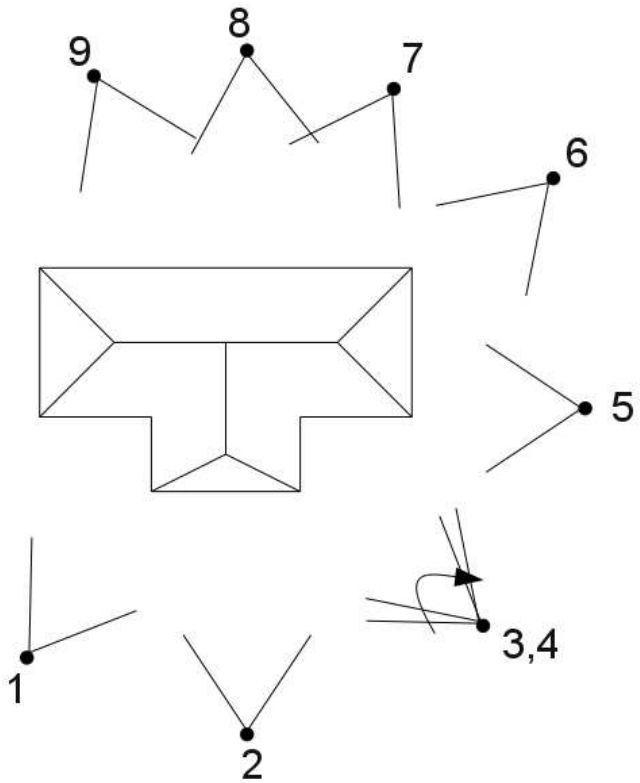


Abbildung 4.3: Bewegungsdegeneration und Strukturdegeneration

Solch degenerierte Daten entstehen in der Praxis häufig und sind entsprechend abzufangen, bevor sie in weitere Berechnungen mit einfließen.
Bewegungsdegenerationen sind schon bei der Erstellung der Aufnahmeserie oder der Selektion der passenden Eingabebilder zu behandeln. Es ist also unbedingt dafür Sorge zu tragen, dass eine hinreichend weite *Baseline* und eine ausreichende Rotation der Blickwinkel zwischen den Aufnahmen vorherrscht.
Um das Problem der degenerierten Strukturen zu lösen, wurde in dieser Arbeit ein manueller Ansatz gewählt. Merkmalspunkte, die in die Schätzung der Fundamental-Matrix einfließen, werden per Hand selektiert und dienen als Eingabedaten des normalisierten 8-Punkte-Algorithmus [HZ03].
Bei der Auswahl dieser Schlüsselpunkte ist also darauf zu achten, dass sie sich weder auf einer gemeinsamen Fläche befinden noch zu sehr auf eine Bildregion konzentriert sind, damit das Ergebnis des 8-Punkte-Algorithmus einer validen Fundamental-Matrix entspricht.

4.1.5 Bestimmung der korrekten Zuordnungen durch geometrische Einschränkungen

Für jedes korrespondierende Punktepaar $\tilde{\boldsymbol{p}}_i^{\mathrm{p}}$ und $\tilde{\boldsymbol{q}}_j^{\mathrm{p}}$ existiert also eine 3 x 3 Matrix vom Rang 2, die *Fundamental-Matrix* $\boldsymbol{F}_{ij}$, sodass gilt:

$$\tilde{\boldsymbol{p}}_i^{\mathrm{p}\mathrm{T}} \boldsymbol{F}_{ij} \tilde{\boldsymbol{q}}_j^{\mathrm{p}} = 0 \tag{4.5}$$

Mit $\tilde{\boldsymbol{p}}_i^{\mathrm{p}} = \begin{bmatrix} x_i & y_i & 1 \end{bmatrix}^{\mathrm{T}}$ und $\tilde{\boldsymbol{q}}_j^{\mathrm{p}} = \begin{bmatrix} x_j & y_j & 1 \end{bmatrix}^{\mathrm{T}}$ die Punkte aus Bild I_i und I_j, $\boldsymbol{F}_{ij}$ die Fundamental-Matrix zwischen den beiden Bildern. Sie bildet einen Punkt $\tilde{\boldsymbol{p}}_i^{\mathrm{p}}$ auf die entsprechende epipolare Linie l in Bild j ab. Gleichung 4.5 ist als die *epipolare Bedingung* bekannt und dient als erster Test der automatischen Zuordnungen auf Korrektheit.
Da in der Praxis die Pixelposition der Bildpunkte einem normalverteilten Rauschen unterliegt und daher nicht ideal ist, wird ein Schwellwert θ eingeführt, innerhalb dessen Grenzen eine Korrespondenz als gültig angenommen wird. θ entspricht somit dem maximal erlaubten Abstand des Punktes $\tilde{\boldsymbol{p}}_j^{\mathrm{p}}$ von der epipolaren Linie l.

$$|\tilde{\boldsymbol{p}}_i^{\mathrm{p}\mathrm{T}} \boldsymbol{F}_{ij} \tilde{\boldsymbol{q}}_j^{\mathrm{p}}| \leq \theta \tag{4.6}$$

Als ein guter Kompromiss zwischen fälschlicherweise aussortierten Punkten und als gültig angesehenen Ausreißern hat sich ein Schwellwert von $\theta = 0.02$ herausgestellt.
Die Fundamental-Matrix bietet demnach ein Mittel, eine mögliche Korrespondenz aus der nächsten Nachbarschaftssuche durch geometrische Bedingungen auf Korrektheit zu prüfen. Sie bezeichnet aber auch die Rotation und Translation zwischen

zwei fortlaufenden Frames auf Pixelebene. Das Pendant der Fundamental-Matrix in normalisierten Bildkoordinaten ist die Essential-Matrix. Sie enthält ebenfalls die Transformation zwischen den Frames und kann dazu genutzt werden, ein erstes Paar an Projektionsmatrizen zu gewinnen (siehe 4.1.6).

4.1.6 Zerlegung der Essential-Matrix

Wie in 4.1.1 beschrieben, wird in dieser Arbeit von einer kalibrierten Kamera ausgegangen. Mittels der aus dem Kalibrierschritt erhaltenen Kalibriermatrix $\boldsymbol{K}$ und der bereits errechneten Fundamental-Matrix $\boldsymbol{F}$ (siehe 4.1.5) lässt sich die Essential-Matrix $\boldsymbol{E}$ wie folgt bestimmen:

$$\boldsymbol{E} = \boldsymbol{K}^{\mathrm{T}}\, \boldsymbol{F}\, \boldsymbol{K} \tag{4.7}$$

Die $\boldsymbol{E}$-Matrix beschreibt die Translation und Rotation zwischen zwei Projektionsmatrizen in Bildkoordinaten. Zur Initialisierung des Modellaufbaus sind die Projektionsmatrizen der ersten beiden Kameras $\boldsymbol{P_1}$ und $\boldsymbol{P_2}$ gesucht. $\boldsymbol{P_1}$ wird so gewählt, dass die Kamera im Weltursprung liegt und ihre optische Achse entlang der Z-Achse des Weltkoordinatensystems ausgerichtet ist:

$$\boldsymbol{P_1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{4.8}$$

Die zweite Projektionsmatrix ergibt sich als diejenige Rotation und Translation, die auf einen Weltpunkt angewendet werden muss, um ihn in das Kamerakoordinatensystem der zweiten Kamera zu transformieren:

$$\tilde{\boldsymbol{p}}^{\mathrm{i}} = \boldsymbol{P_2}\, \tilde{\boldsymbol{p}}^{\mathrm{w}} \tag{4.9}$$

und

$$\boldsymbol{P_2} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \end{bmatrix} \tag{4.10}$$

Die $\boldsymbol{E}$-Matrix setzt sich zusammen als das Kreuzprodukt der gesuchten Translation und der Rotation zwischen zwei Projektionsmatrizen:

$$\boldsymbol{E} = [\boldsymbol{t}]_{\times}\, \boldsymbol{R} \tag{4.11}$$

$[\boldsymbol{t}]_{\times}$ ist dabei die Matrixdarstellung eines Kreuzprodukts aus $\boldsymbol{R}$ und $\boldsymbol{t}$. Diese Tatsache wird dazu genutzt, durch eine geeignete Faktorisierung der Essential-Matrix an die benötigte Information der Rotation und Translation zu gelangen. [HZ03] zeigt dazu eine Vorgehensweise, die im Folgenden ausgeführt ist.

Die E-Matrix wird in eine schiefsymmetrische Matrix $\boldsymbol{S}$ und eine Rotationsmatrix $\boldsymbol{R}$ zerlegt:

$$\boldsymbol{E} = [\boldsymbol{t}]_\times \boldsymbol{R} = \boldsymbol{S}\,\boldsymbol{R} \tag{4.12}$$

Als Hilfsmatrizen werden definiert:

$$\boldsymbol{W} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{Z} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Und die Singulärwertzerlegung von $\boldsymbol{E}$ ist:

$$\boldsymbol{U} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{V}^{\mathrm{T}} = SVD(\boldsymbol{E}) \tag{4.13}$$

Dadurch ergeben sich folgende mögliche Zerlegungen für $\boldsymbol{E} = \boldsymbol{S}\,\boldsymbol{R}$:

$$\begin{gathered} \boldsymbol{S} = \boldsymbol{U}\,\boldsymbol{Z}\,\boldsymbol{U}^{\mathrm{T}} \\ \boldsymbol{R_1} = \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^{\mathrm{T}} \\ \boldsymbol{R_2} = \boldsymbol{U}\,\boldsymbol{W}^{\mathrm{T}}\,\boldsymbol{V}^{\mathrm{T}} \end{gathered}$$

Darüber hinaus gilt:

$$\boldsymbol{t}^{\mathrm{T}}\,\boldsymbol{E} = \boldsymbol{t}^{\mathrm{T}}[\boldsymbol{t}]_\times \boldsymbol{R} = 0 \tag{4.14}$$

Die gesuchte Translation liegt also im linken Nullraum von $\boldsymbol{E}$ und lässt sich direkt als letzter Spaltenvektor von $\boldsymbol{U}$ (im Folgenden als $\boldsymbol{u_3}$ bezeichnet) bestimmen.

Das Vorzeichen der Essential-Matrix lässt sich jedoch nicht durch deren Zerlegung ermitteln. Unter der Annahme $\boldsymbol{P_1} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}$ ergeben sich für $\boldsymbol{P_2}$ vier mögliche Konfigurationen:

$$\begin{gathered} \begin{bmatrix} \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^{\mathrm{T}} & +\boldsymbol{u_3} \end{bmatrix}, \\ \begin{bmatrix} \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^{\mathrm{T}} & -\boldsymbol{u_3} \end{bmatrix}, \\ \begin{bmatrix} \boldsymbol{U}\,\boldsymbol{W}^{\mathrm{T}}\,\boldsymbol{V}^{\mathrm{T}} & +\boldsymbol{u_3} \end{bmatrix}, \\ \begin{bmatrix} \boldsymbol{U}\,\boldsymbol{W}^{\mathrm{T}}\,\boldsymbol{V}^{\mathrm{T}} & -\boldsymbol{u_3} \end{bmatrix} \end{gathered}$$

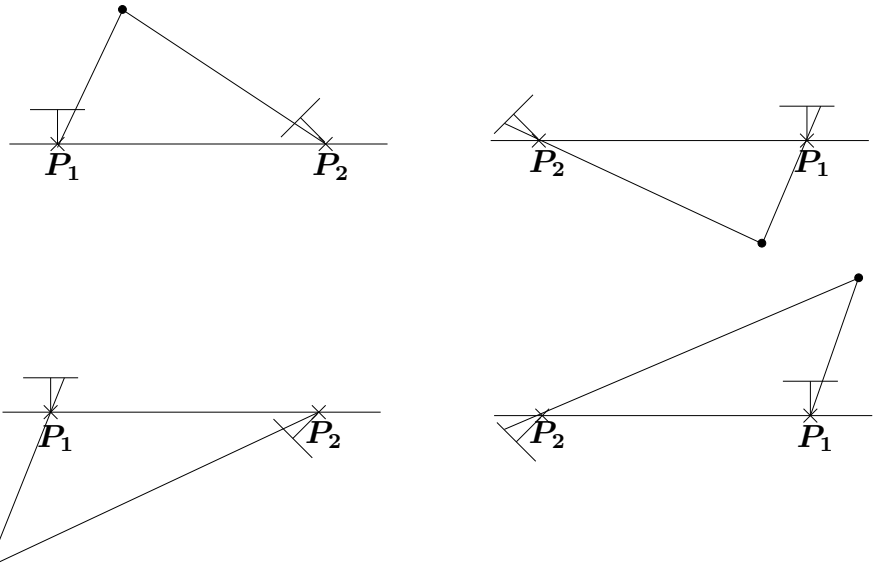


Abbildung 4.4: Die vier möglichen Zusammensetzungen der zweiten Projektionsmatrix, Quelle: [HZ03]

Abbildung 4.1.6 veranschaulicht die vier Alternativen geometrisch. Die jeweils linke Spalte entspricht der positiven Translation der Baseline, die jeweils rechte Spalte einer negativen Bewegung. Die Zeilen der Darstellung unterscheiden sich in einer um 180° um die Baseline gedrehten Kamera.

Die einzig gültige Zusammensetzung für $\boldsymbol{P_2}$ ergibt sich aus der Tatsache, dass ein triangulierter Punkt vor beiden Kamerazentren liegen muss. Zur Auflösung dieser Mehrdeutigkeit wird folglich ein beliebiger Punkt mit allen vier Alternativen trianguliert und jeweils dessen Tiefenwerte aus Sicht beider Kameras untersucht. Dieser Punkt wird hier als der erste Weltpunkt der automatischen Zuordnungen angenommen, unter der Bedingung, dass sie die epipolare Bedingung erfüllen (siehe 4.1.5).
Das Ergebnis der E-Matrix-Zerlegung ist die Projektionsmatrix $\boldsymbol{P_2}$ im Kamerakoordinatensystem von $\boldsymbol{P_1}$.

4.1.7 Einreihung weiterer Bilder in das Modell

Auf Basis der E-Matrix-Zerlegung lassen sich neben den ersten beiden Kameras auch weitere sequentiell eingereihte Kameras berechnen.
Die Translation zwischen zwei Projektionsmatrizen ist nur in ihrer Richtung eindeutig, nicht aber in ihrem Betrag. Es besteht jedoch ein direkter linearer Zusammenhang zwischen den Abständen der Weltpunkte des Modells und der Größe der

Translation der zweiten Projektionsmatrix. Ein um einen Skalierungsfaktor s skaliertes Modell erscheint in seiner Projektion im Kamerabild der zweiten Kamera genauso groß wie ein kleiner skaliertes Modell bei entsprechend geringerem Translationsbetrag.
Diese Tatsache kann dazu verwendet werden, die Translation zwischen den Frames entsprechend konsistent mit den Abständen der Weltpunkte zu halten und somit ein korrekt skaliertes Modell zu generieren. Um einen entsprechenden Skalierungsfaktor zu errechnen werden also folgende Informationen benötigt: $\boldsymbol{P_{i-1}}$, die Projektionsmatrix des Frames F_{i-1}, $\boldsymbol{P_i}$ und $\boldsymbol{P_{i+1}}$ die beiden Projektionsmatrizen der entsprechenden Frames F_i und F_{i+1}. Außerdem die Korrespondenzen der Bilder $i-1$ und i sowie i und $i+1$.
Über eine Verkettung der Korrespondenzen von drei aufeinanderfolgenden Frames lassen sich diejenigen Merkmale ermitteln, die in allen drei Bildern sichtbar sind (siehe dazu Abbildung 4.6 und Abschnitt 4.1.9).
Sie werden nun aus dem schon korrekt skalierten Projektionsmatrizenpaar $\boldsymbol{P_{i-1}}$ und $\boldsymbol{P_i}$ und dem neuen Projektionsmatrizenpaar $\boldsymbol{P_i}$ und $\boldsymbol{P_{i+1}}$ in Weltpunkte trianguliert (siehe 4.1.8). Pro Framepaar werden daraufhin die Abstände zweier solcher errechneter Weltpunkte bestimmt. Dieser Vorgang wird für alle Merkmale, die in den drei Ansichten vorhanden sind, durchgeführt und deren Abstände jeweils ins Verhältnis gesetzt.
Ein möglicher Skalierungsfaktor der Translation von $\boldsymbol{P_{i+1}}$ ergibt sich aus dem Median der Abstandsverhältnisse der jeweils verketteten Weltpunkte. Auch der Mittelwert dieser Abstände ist ein möglicher Skalierungsfaktor, er liefert aber bis auf eine zu vernachlässigende Abweichung von etwa 0.01 denselben Wert.
Wie in 4.1.6 bereits erwähnt wird bei der E-Matrix-Zerlegung die jeweils erste Projektionsmatrix als Einheitsmatrix in der Rotation und einem Nulltranslationsvektor angenommen. Werden sequentiell weitere Projektionsmatrizen in das Modell eingespeist, muss dafür Sorge getragen werden, dass $\boldsymbol{P_i}$ zunächst in das Koordinatensystem von $\boldsymbol{P_{i-1}}$ transformiert und daraufhin die Rotation und Translation aus $\boldsymbol{P_i}$ aufgerechnet wird. Durch eine E-Matrix-Zerlegung wird die *relative Orientierung* der beiden Kameras beschrieben. Das bedeutet, dass ein möglicher Fehler, der bei der E-Matrix-Zerlegung oder der Bestimmung des korrekten Skalierungsfaktors entsteht, über das gesamte Modell getragen wird. Abbildung 4.5 veranschaulicht die Auswirkung eines nicht konsistenten Skalierungsfaktors. Die Szene zeigt eine eigentlich planare Fläche, deren Weltpunkte im Modell versetzt trianguliert wurden. Dieser Fehler ist auf eine unkorrekte Kameraposition durch einen falschen Skalierungsfaktor ihres Translationsvektors zurückzuführen. Eine solche Fehlschätzung fließt dann in alle folgenden Projektionsmatrizen mit ein und akkumuliert sich bei weiteren inkonsistenten Werten.

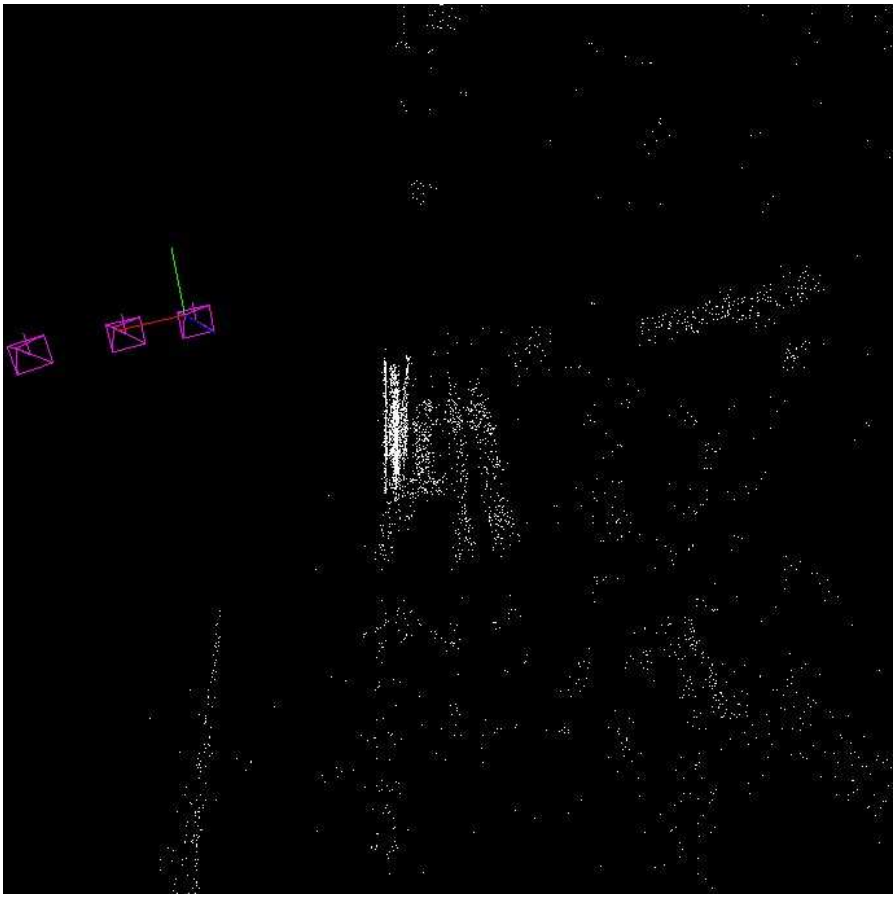


Abbildung 4.5: Anfügen weiterer Bilder durch E-Matrix-Zerlegung

Eine Lösung dieses Problems wird in 4.1.9 beschrieben. Die Projektionsmatrizen werden in ihrer *absoluten Orientierung* in Bezug auf das Weltkoordinatensystem bestimmt. Es werden bereits hier die 3D-2D-Korrespondenzen zur Berechnung der Kameramatrizen verwendet, die E-Matrix-Zerlegung dient nur noch zur Initialisierung des Modells.

4.1.8 Triangulierung der Weltpunkte

Nun stehen alle Informationen bereit, die für die Rückrechnung der Weltposition, die durch die jeweiligen Bildmerkmale beschrieben ist, benötigt werden.
Ziel der Modellerstellung ist die Herstellung des Zusammenhangs zwischen einem Bildpunkt und seinem dreidimensionalen Ursprung. Ein Bild kann als projektiver Raum $\mathbb{P}^2$ angesehen werden und bildet sich aus der Projektion des dreidimensio-

nalen Raumes $\mathbb{P}^3$ auf jene Bildebene.
Bei bekannten Projektionsmatrizen zweier Kameras $\boldsymbol{P_1}$ und $\boldsymbol{P_2}$ und den korrespondierenden Bildpunkten $\tilde{\boldsymbol{p}}^{\mathrm{i}}$ und $\tilde{\boldsymbol{q}}^{\mathrm{i}}$, die der Projektion eines Weltpunktes $\tilde{\boldsymbol{w}}^{\mathrm{w}}$ in deren Bildebenen entsprechen, lässt sich die Position des Weltpunktes schätzen als:

$$\tilde{\boldsymbol{w}}^{\mathrm{w}} = \underset{\tilde{\boldsymbol{w}}^{\mathrm{w}}}{\operatorname{argmin}} \left\| \boldsymbol{P_1}\, \tilde{\boldsymbol{w}}^{\mathrm{w}} - \tilde{\boldsymbol{p}}^{\mathrm{i}} \right\|^2 \cdot \left\| \boldsymbol{P_2}\, \tilde{\boldsymbol{w}}^{\mathrm{w}} - \tilde{\boldsymbol{q}}^{\mathrm{i}} \right\|^2 \tag{4.15}$$

Der gesuchte Weltpunkt $\tilde{\boldsymbol{w}}^{\mathrm{w}} \in \mathbb{P}^3$ ist derjenige Punkt, dessen quadratischer Abstand seiner Projektion in beide Bilder und der betrachteten Bildpunkte minimal ist. Diese Darstellung entspricht der Minimierung des Rückprojektionsfehlers des gesuchten Weltpunktes in beide Bilder. Nach der Umformung von 4.15 in ein homogenes Gleichungssystem und dessen Lösung durch Singulärwertzerlegung liegt $\tilde{\boldsymbol{w}}^{\mathrm{w}}$ in homogenen Weltkoordinaten vor:

$$\tilde{\boldsymbol{w}}^{\mathrm{w}} = \begin{bmatrix} x^{\mathrm{w}} & y^{\mathrm{w}} & z^{\mathrm{w}} & 1 \end{bmatrix}^{\mathrm{T}} \tag{4.16}$$

Als letzter Test, bevor der triangulierte Punkt und dessen Bildpunkte als gültig angesehen werden, ist zu überprüfen, ob seine Position auch tatsächlich im Sichtfeld beider Kameras liegt.
Dazu wird der Weltpunkt in das entsprechende Kamerakoordinatensystem überführt:

$$\tilde{\boldsymbol{w}}^{\mathrm{c}} = \boldsymbol{P_i} \tilde{\boldsymbol{w}}^{\mathrm{w}} \tag{4.17}$$

mit $\tilde{\boldsymbol{w}}^{\mathrm{c}} = \begin{bmatrix} x^{\mathrm{c}} & y^{\mathrm{c}} & z^{\mathrm{c}} & 1 \end{bmatrix}^{\mathrm{T}}$. Bei bestandenem Tiefentest, das heißt $z^{\mathrm{c}} > 0$, wird der Weltpunkt und seine zweidimensionalen Bildpunkte als valide anerkannt.

Zwei Bildpunkte, die alle bisherigen Bedingungen erfüllt haben gelten fortan als korrespondierend.

4.1.9 2D-3D-Zuordnung

Bisher wurde eine Korrespondenz sehr unformell beschrieben und es ist nur bekannt, wann zwei Bildmerkmale p und q korrespondieren: sie sind nächste Nachbarn, haben eine $distanceRatio < r$, erfüllen die epipolare Bedingung und deren triangulierter Weltpunkt liegt vor beiden Kamerazentren. Diese Eigenschaft der Zusammengehörigkeit lässt sich aber auch als Relation auf der Menge der gesamten Merkmale M betrachten:

$$\mathrm{correspond}(p, q) \subseteq M \times M \tag{4.18}$$

Die Korrespondenzrelation hat wichtige Eigenschaften.
Sie ist injektiv:

$$\forall p \in M_1, q_1 \in M_2 : \text{correspond}(p, q_1) \Rightarrow \nexists q_2 \in M_2 : \text{correspond}(p, q_2) \tag{4.19}$$

Sie ist symmetrisch:

$$\forall p, q \in M : \text{correspond}(p, q) \Rightarrow \text{correspond}(q, p) \tag{4.20}$$

Bei der Korrespondenzbestimmung der Merkmale zwischen zwei Bildern ist die Reihenfolge der Bilder also nicht ausschlaggebend.

Sie ist transitiv:

$$\forall p, q, z \in M : \text{correspond}(p, q) \wedge \text{correspond}(q, z) \Rightarrow \text{correspond}(p, z) \tag{4.21}$$

Korrespondieren Merkmalen zweier Frames F_{i-1} und F_i (wobei p ein Merkmal aus F_{i-1} und q ein Merkmal aus F_i ist) und existiert eine Zuordnung zwischen F_i und F_{i+1} mit q und z aus eben jenen fortlaufenden Frames, so besteht implizit auch die Korrespondenz zwischen p und z.

Dieser Zusammenhang ist an zwei Stellen der Arbeit besonders von Interesse. Die Initialisierung des Modells erfolgt wie in 4.1.6 beschrieben durch eine Zerlegung der Essential-Matrix der ersten beiden Frames. Die Korrespondenzen zwischen diesem Bildpaar werden anschließend in Weltpunkte trianguliert (siehe 4.1.8) und die gewonnene Information der Tiefe jedem Bildpunkt zugeordnet. Anhand der Transitivitätseigenschaften der Korrespondenzrelation lassen sich nun die Projektionsmatrizen fortlaufender Frames durch 2D-3D-Zuordnungen bestimmen. Die Problematik der inkonsistenten Skalierungsfaktoren der Translation (siehe 4.1.7) ist gelöst, da die so geschätzten Kameramatrizen direkt in Weltkoordinaten definiert sind und nicht auf den vorherigen Kameramatrizen beruhen. Abbildung 4.6 zeigt eine Verkettung über drei Bilder: die bereits triangulierten Merkmale (aus F_{i-1} und F_i) und deren Bildzuordnungen des rechten Frames sind die Basis für die Bestimmung der Projektionsmatrix des Frames F_{i+1}. Auf die Berechnung einer Projektionsmatrix anhand von 2D-3D-Korrespondenzen wird in 4.2.4 und 4.2.6 eingegangen.

Die zweite Anwendung dieser Verhältnisse ist in der Erstellung des Modells selbst zu sehen. Ein Modell muss die Information bereitstellen, die für die Schätzung der Kamerapose benötigt wird. Dies ist eben genau jener Zusammenhang der Weltpunkte und ihrer Bildpunkte. Über den Vergleich der Deskriptorvektoren und diese Zuordnung lassen sich bei der Posebestimmung Rückschlüsse über die Position eines Bildpunktes im dreidimensionalen Raum gewinnen.

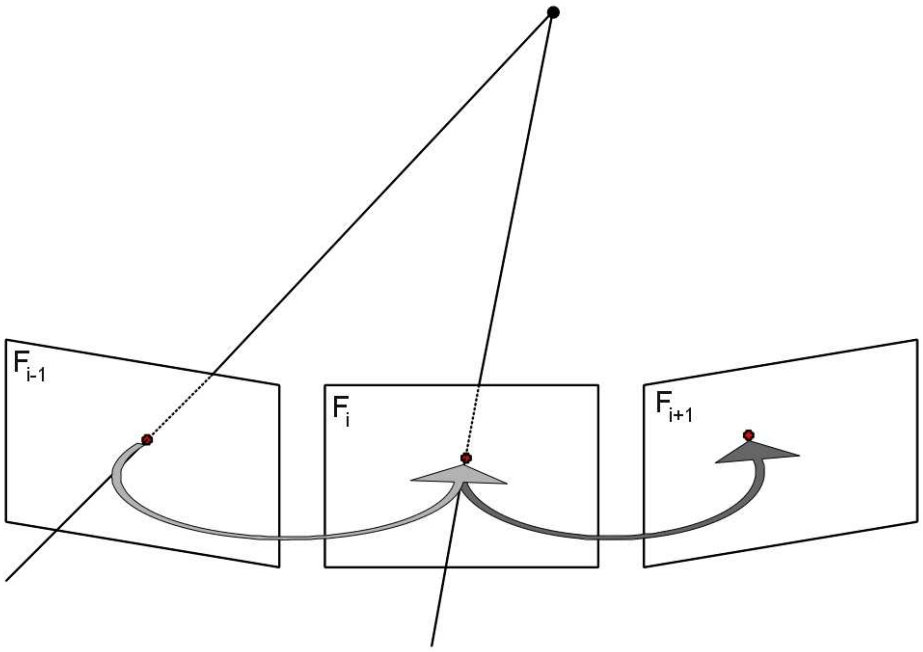


Abbildung 4.6: Ermittlung der 2D-3D-Zuordnung über eine fortlaufende Bildreihe

Ein Weltpunkt ist physisch ein Punkt, der in mehreren Bildern durch ein Bildmerkmal beschrieben ist. Soll die mehrfache Beschreibung durch mehrfache Deskriptorvektoren an einen Weltpunkt annotiert werden, erweisen sich 4.20 und 4.21 als nützlich. Wegen der Symmetrie lässt sich ein Bildpunkt über seine Korrespondenzen auch rückwärts verfolgen. Besteht bereits eine Verbindung über drei Ansichten und ist ein Weltpunkt aus den ersten Ansichten errechnet, wird an diesen die neue Information aus dem aktuellen Bild angefügt. Dadurch vergrößert sich der Beschreibungsradius des Weltpunktes mit jeder neuen Ansicht. Das ist bei der späteren Posebestimmung von enormen Nutzen.

Bei der Erstellung des Modells wird diese Beziehung für jedes sequentiell neu hinzugefügte Bild untersucht. Ist für ein Bildmerkmal noch kein Weltpunkt beschrieben, wird ein neuer Weltpunkt erstellt. Besteht bereits eine Korrespondenz, so wird sie aktualisiert.

Für jeden Weltpunkt $\tilde{\boldsymbol{p}}^{\mathrm{w}}$ ist nach Erstellung des Modells folgende Information verfügbar:

- Seine Koordinaten im Weltkoordinatensystem
- Die Deskriptoren sämtlicher korrespondierender Merkmalspunkte der Bildserie, inklusive derer Bildposition, Orientierung und Skala
- Zu jedem Deskriptor der Index des Bildes, in dem das Merkmal detektiert wurde
- Die zugehörige Projektionsmatrix dieses Bildes

Die 2D-3D-Zusammengehörigkeit beschreibt die Eigenschaften eines Modells. Für eine Menge an Bildpunkten $\tilde{\boldsymbol{p}}^{\mathrm{i}} \in \mathbb{P}^2$, eine Menge an Weltpunkten $\tilde{\boldsymbol{w}}^{\mathrm{w}} \in \mathbb{P}^3$ und den Projektionsmatrizen $\boldsymbol{P}$ ist ein Modell definiert:

$$\begin{aligned} \mathrm{model}(\tilde{\boldsymbol{p}}^{\mathrm{i}}, \tilde{\boldsymbol{w}}^{\mathrm{w}}, \boldsymbol{P}) \Leftrightarrow (\forall \tilde{\boldsymbol{w}}^{\mathrm{w}} \in \mathbb{P}^3 : \exists \tilde{\boldsymbol{p}}^{\mathrm{i}} \in \mathbb{P}^2 \mid \tilde{\boldsymbol{p}}^{\mathrm{i}} = \boldsymbol{P_i} \tilde{\boldsymbol{w}}^{\mathrm{w}}) \\ \wedge (\forall \tilde{\boldsymbol{p}}^{\mathrm{i}} \in \mathbb{P}^2, \tilde{\boldsymbol{q}}^{\mathrm{i}} \in \mathbb{P}^2, \tilde{\boldsymbol{w}}^{\mathrm{w}} \in \mathbb{P}^3 : \\ \mathrm{correspond}(\tilde{\boldsymbol{p}}^{\mathrm{i}}, \tilde{\boldsymbol{q}}^{\mathrm{i}}) \Rightarrow \exists! \tilde{\boldsymbol{w}}^{\mathrm{w}} \mid \tilde{\boldsymbol{p}}^{\mathrm{i}} = \boldsymbol{P_i} \tilde{\boldsymbol{w}}^{\mathrm{w}} \wedge \tilde{\boldsymbol{q}}^{\mathrm{i}} = \boldsymbol{P_j} \tilde{\boldsymbol{w}}^{\mathrm{w}}) \end{aligned} \tag{4.22}$$

mit $i \neq j$ den Indices der Bilder und derer Projektionsmatrizen.
Über einen Vergleich der Deskriptoren des Modells und der Deskriptoren eines neuen Bildes, das Teilbereiche des Modells als Inhalt hat, kann somit die dreidimensionale Position der Merkmale bestimmt werden. Die Beziehung bildet die Grundlage für die Bestimmung der Kamerapose eines Bildes. In 4.2 werden die Schritte für die letztendliche Bestimmung der Rotations- und Translationsparameter erläutert.

4.2 Schätzung der Kamerapose

Die Schätzung der Kamerapose ist in dem Problem der Ermittlung der externen Orientierung einer Kamera zu sehen, ohne ein Vorab-Wissen über deren Lage im Raum. Sie basiert lediglich auf den Korrespondenzen zwischen Bildmerkmalen (2D) und deren Weltpunkte (3D).
Der Ablauf der Posebestimmung ist in die Schritte gegliedert:

- Extraktion der Merkmale des Posebildes
- Zusammenfassung aller Modellmerkmale in eine globale Liste
- Matching der Posemerkmale gegen alle Modellmerkmale
- Aussortieren mehrfacher Merkmalsbelegung einer Position im Posebild

- Robuste Schätzung der Kamerapose anhand von 2D-3D-Korrespondenzen

Abbildung 4.7 veranschaulicht den Ablauf in einem Aktivitätsdiagramm. Ein Bild, das nicht aus der Serie stammt, die bei der Modellerstellung zum Einsatz kam, wird fortan als *Posebild* bezeichnet. Analog dazu die aus dem Posebild extrahierten SIFT-Merkmale als *Posemerkmale*. Sämtliche Merkmale des Modells werden in einer *globalen Keyliste* zusammengefasst.

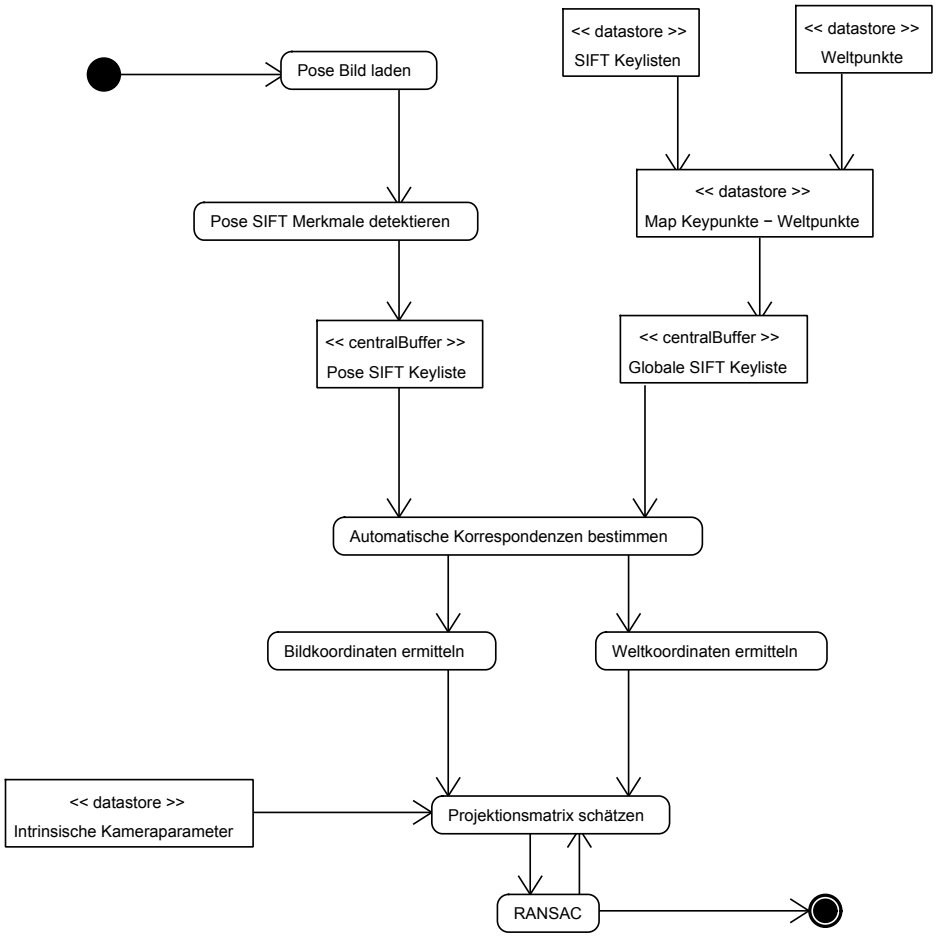


Abbildung 4.7: Aktivitätsdiagramm der Poseschätzung

4.2.1 Korrespondenzbestimmung zwischen Referenzbild und Modell

Abbildung 3.4 zeigt eine Merkmalsliste, wie sie für jedes Eingabebild der Modellserie existiert. Sämtliche Merkmale, die aus den konsekutiven Aufnahmen der Eingabebilder in das Modell eingeflossen sind, werden zunächst in eine globale Liste zusammengefasst.
Anhand dieser globalen Liste und den neu extrahierten Posemerkmalen werden in einem automatischen Zuordnungsvorgang die möglichen Kandidaten für eine Poseschätzung ermittelt. Dabei kommt erneut das in 4.1.3 erläuterte Distanzverhältnis eines nächsten Nachbarn zu seinem zweitnächsten Nachbarn zum Einsatz. Merkmalszuordnungen, die für die Posebestimmung ungeeignet sind, werden so vorab aussortiert.
Die verbleibenden Zuordnungen sind die Grundlage des weiteren Vorgehens: ein nächster Nachbar der Modellmerkmale zu einem gegebenen Posemerkmal hat aus den Eigenschaften des Modells (siehe 4.22) genau eine Weltposition zugeordnet. Demnach wird diese Weltposition auch für das entsprechende Posemerkmal angenommen. Sind genügend Posemerkmale mit diesen Eigenschaften bestimmt, kann die Orientierung und Lage der Kamera ermittelt werden.

4.2.2 Eliminierung mehrfacher lokaler Merkmalsbelegung

Der SIFT-Algorithmus detektiert an einigen lokalen Positionen mehrfach Merkmale. Wie in 3.1.4 beschrieben, wird zu jeder weiteren dominanten Orientierung, die innerhalb von 80% des Spitzenwertes der Orientierungen liegt, ein neues Merkmal erstellt. Diese Merkmale liegen an derselben Bildposition, haben aber unterschiedliche Hauptorientierungen Θ. Das Prinzip trägt zwar wesentlich zur Stabilität der Scale Invariant Feature Transform bei [Low04], ist jedoch im Falle der Poseschätzung von Nachteil.
Die Schätzung der Rotation und Translation einer Kamera erfolgt durch lineare Optimierungsverfahren anhand von 2D-3D-Korrespondenzen. Werden alle Punkte, die durch ihre Nachbarschaft potentiell einem Punkt des Modells entsprechen in die Schätzung der Kamera miteinbezogen, so werden diese mehrfach belegten Bildpositionen auch mehrfach gewichtet, ohne zusätzlich nützliche Information mit einzubringen.
Ist die Anzahl der Korrespondenzen zwischen Modell und Posebild nahe an der minimalen Grenze, die zur Bestimmung der Modellparameter nötig ist und treten solche Mehrfachbelegungen auf, ist die gelieferte Information nur scheinbar ausreichend, ein gültiges Gleichungssystem aufzustellen.
Deshalb sind Bildmerkmale, die sowohl im Bildraum als auch deren Zuordnung im

dreidimensionalen Raum sehr nahe zusammenliegen, nur einfach in die Schätzung aufzunehmen.

4.2.3 Robuste Schätzung der Kamerapose

Wie bereits mehrfach erwähnt ist die automatische Korrespondenzbestimmung zwischen Merkmalen aufgrund von Fehlzuordnungen, die nicht als solche erkannt werden, ausreißerbehaftet. Die Korrektheit des Ergebnisses linearer Optimierungsverfahren ist aber stark abhängig von der Korrektheit der Eingabedaten. Kleine Ausreißer in der Messmenge führen zu großen Ergebnisschwankungen. Diese Art der Verfahren werden deshalb als *nicht robust* bezeichnet.
Im Zuge der Modellerstellung standen noch geometrische Mittel zur Verfügung, die automatischen Zuordnungen auf ihre Korrektheit zu prüfen (siehe 4.1.5 und 4.1.8). Da aber weder die Essential-Matrix noch die Projektionsmatrix zum Zeitpunkt der Poseschätzung bekannt ist, muss ein anderes Verfahren zur Erhöhung der Robustheit eingesetzt werden.

Ein solches generelles Verfahren, um mit ausreißerbehafteten Daten umzugehen, ist der RANdom SAmple Consensus (RANSAC) Algorithmus, der bereits 1981 in [FB81] präsentiert wurde.
RANSAC ist in der Lage in eine fehlerhafte Datenmenge korrekte Modelle einzupassen. Dazu wird zufällig die minimale Anzahl an Daten gezogen, die zur Schätzung der Modellparameter (also hier die Parameter der Projektionsmatrix) benötigt wird. Mit ihnen wird ein hypothetisches Modell erstellt und die Qualität des Modells anhand der Menge der Daten bemessen, die das Modell unterstützen. Die unterstützenden Punkte werden *Support Set* oder *Consensus Set* genannt und werden anhand eines Schwellwerts τ überprüft. Im Falle der Projektionsmatrix ist dieser Schwellwert der maximal erlaubte Rückprojektionsfehler des Weltpunktes in den Bildpunkt mit der hypothetisch erstellten Projektionsmatrix.
Dieser Vorgang wird iterativ fortgeführt, bis sich ein genügend großes Consensus Set gefunden hat und das Modell als gut bewertet wird.
Die maximale Anzahl an RANSAC-Iterationen N ist gegeben durch:

$$N = \frac{\log(1-p)}{\log(1-w^n)} \tag{4.23}$$

Bei einer Wahrscheinlichkeit von p, dass mindestens eine gezogene Auswahl ein ausreißerfreier Satz von Punkten ist, einem geschätzten Inlierverhältnis von w in der Messmenge und der Anzahl an benötigten Modellparametern n.
Die initiale Belegung dieser Werte ist in Tabelle 4.1 aufgezeigt. Beide Ansätze der Poseschätzung, die in dieser Arbeit verwendet werden, haben eine minimal

benötigte Anzahl von $n = 6$ 2D-3D-Korrespondenzen (siehe 4.2.4 und 4.2.6).

Parameter	Wert
p	0,99
τ	0,005
w	0,2
n	6

Tabelle 4.1: Initiale Belegung der RANSAC-Parameter

Diese Belegungen sind empirisch ermittelt worden, können aber während der Laufzeit des Programms angepasst werden (siehe C.1.2).

Ist der RANSAC-Durchlauf erfolgreich abgeschlossen und das Support Set (also die Inlier) bestimmt, folgt mittels allen Inliern die letztendliche Schätzung der Kamerapose. In den folgenden Abschnitten 4.2.4 und 4.2.6 werden dazu zwei Varianten, die hier zum Einsatz kommen, erläutert.

4.2.4 Direkte Lineare Transformation

Eine erste Methode, die zur Schätzung der Projektionsparameter verwendet wird ist die Direkte Lineare Transformation, auch als DLT bekannt. Sie ist auf vielfältigen Anwendungsbereichen der 3D-Rekonstruktion einsetzbar.
Der grundlegende Gedanke der DLT ist das Kollinearitätsprinzip: die Projektion eines Weltpunktes in ein Bild liegt auf der Geraden des Kameraursprungs zu jenem Weltpunkt. Diese Tatsache kann dazu genutzt werden ein überbestimmtes lineares Gleichungssystem aufzustellen. Eine Mindestanzahl von $n = 6$ Bildpunkt-Weltpunkt-Korrespondenzen ist im Falle der 12 unbekannten Parameter zur Lösung des Gleichungssystems nötig (siehe B.1). Nach einer Umsortierung der Parameter wird der Lösungsraum mittels Singulärwertzerlegung bestimmt. Neben der trivialen Lösung entsteht somit eine eindeutige Lösung der Rotations- und Translationsparameter der Kamera.

4.2.5 Korrektur der Rotationsparameter

Es ist bekannt, dass eine Rotationsmatrix lediglich drei Freiheitsgrade (DOF) in ihrer Bewegung hat. Die DLT schätzt die 9 Parameter der Rotation jedoch völlig unabhängig voneinander, ohne die geforderten Bedingungen an eine Rotationsmatrix zu berücksichtigen. Eine Rotationsmatrix ist orthonormal und muss die Orien-

tierung der Referenzkoordinatensysteme stets erhalten (siehe B.2). Sind trotz allen bisher vorgenommenen Überprüfungen noch Ausreißer in der Messdatenmenge, so werden bei der direkten linearen Transformation diese Bedingungen ignoriert.
Das bedeutet in der Konsequenz, dass eine von der DLT gelieferte Rotationsmatrix nicht per se gültig ist.
Die Eigenschaften einer Rotationsmatrix lassen sich aber mit Hilfe der Singulärwertzerlegung erzwingen. In B.2 ist ein Vorgehen aufgezeigt, das die geforderten Eigenschaften einer echten Rotationsmatrix wiederherstellt. Das Resultat ist die nächste Rotationsmatrix zu der aus der DLT errechneten Quasi-Rotationsmatrix. Der Begriff *nächste* ist hier im Sinne der Frobenius-Norm der resultierenden Matrix zu verstehen.
Wird die so korrigierte Rotationsmatrix als Rotation einer Projektionsmatrix angenommen, resultiert dies in ähnlichen Effekten, die schon bei der E-Matrix-Zerlegung beobachtet wurden (siehe Abbildung 4.5). Eine im Nachhinein korrigierte Projektionsmatrix ist zwangsläufig nicht mehr konsistent mit den Eingabedaten, die zur Ermittlung der ursprünglichen Rotationsparameter verwendet wurden. Wird die korrigierte Projektionsmatrix während der Modellerstellung für eine Triangulierung der Weltpunkte verwendet, so entsteht ein Rückprojektionsfehler.
Derselbe theoretische Fehler entsteht bei der Schätzung der Poseparameter.
Die DLT ist somit nur im optimalen, rauschfreien Fall ein probates Mittel zur Schätzung einer Kameramatrix.

4.2.6 Lineare Optimierung nach Fiore

Ein Verfahren, das die Orthonormalitätsbedingung einer Rotation gewährleistet, ist das Verfahren nach Fiore [Fio01].
Das Problem der externen Orientierung der Kamera ist in seiner vollständigen Form gegeben als:

$$l_i \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = s\boldsymbol{R}(a_i + \boldsymbol{t}) \tag{4.24}$$

mit l_i die Tiefe des Bildpunktes $\begin{bmatrix} x_i & y_i & 1 \end{bmatrix}^{\mathrm{T}}$ aus Kamerasicht, der Skalierung s der Rotation und Translation und dem Weltpunkt a_i.
Die DLT beruht lediglich auf dem Kollinearitätsprinzip der Bild- und Weltpunkte, die tatsächlichen Tiefenwerte der Weltpunkte sind dabei nicht von Interesse. Wie jedoch in 4.1.7 bereits erwähnt, besteht ein direkter Zusammenhang zwischen der Entfernung der Weltpunkte zur Kamera und der Skalierung des Modells (respektive der Skalierung der Rotation und Translation).
Bei der Methode nach Fiore wird dieser Zusammenhang dazu verwendet, zunächst

die Tiefenwerte l_i der Punkte zu extrahieren, den Skalierungsfaktor s zu errechnen und daraufhin die gesuchte Rotation und Translation zu bestimmen.
Die l_i-Parameter aus 4.24 werden durch eine geschickt gewählte Gewichtungsmatrix $\boldsymbol{W}$ extrahiert, sodass gilt:

$$\sum_{i=1}^{N} w_{ij} l_i \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \sum_{i=1}^{N} w_{ij} s\boldsymbol{R}(a_i + \boldsymbol{t}) = 0 \tag{4.25}$$

für alle $j = 1, ..., N - 4$.
Die Matrix $\boldsymbol{W}$ setzt sich aus den Singulärvektoren korrespondierend zum rechten Nullraum der Weltpunktmatrix $\boldsymbol{X}$ (einer Matrix, die aus dem spaltenweisen Anfügen der Weltpunkte a_i besteht) zusammen:

$$\boldsymbol{X}\boldsymbol{W} = 0 \tag{4.26}$$

Durch Umsortierung folgt, dass der Lösungsvektor l (bestehend aus den Einträgen von l_i) im linken Nullraum von $\boldsymbol{W}$ liegt. Dieser wird nach 4.26 von $\boldsymbol{X}^{\mathrm{T}}$ aufgespannt und somit ist l bis auf einen gemeinsamen unbekannten Faktor γ bestimmt durch:

$$l = \boldsymbol{X}^{\mathrm{T}}\gamma \tag{4.27}$$

Die Summe der w_i Elemente, die Bildpunkte ohne deren homogene Koordinate und $\boldsymbol{X}^{\mathrm{T}}$ werden in eine Matrix $\boldsymbol{C}$ zusammengefasst:

$$\boldsymbol{C} = \sum_{i=0}^{N} w_{ij} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \boldsymbol{X}^{\mathrm{T}} \tag{4.28}$$

für alle $j = 1, ..., N - 4$.
Der vordere Teil aus 4.25 kann nun umformuliert werden:

$$\boldsymbol{C}\gamma = 0 \tag{4.29}$$

Der gemeinsame Faktor γ lässt sich durch Nullraumbestimmung von $\boldsymbol{C}$ lösen und die Tiefenparameter l durch 4.27 bestimmen.

Der linke Teil aus 4.24 ist nun eindeutig beschrieben und wird fortan als b_i bezeichnet:

$$b_i = s\boldsymbol{R}(a_i + \boldsymbol{t})\forall i \tag{4.30}$$

Zur Lösung der restlichen Unbekannten $(s, \boldsymbol{R}, \boldsymbol{t})$ werden die Schwerpunkte a_0 und b_0 der beiden Punktmengen a_i und b_i errechnet. Außerdem die Distanzen

$\tilde{a}_i = a_i - a_0$ und $\tilde{b}_i = b_i - b_0$.
Der Skalierungsfaktor s ist gegeben durch die Minimierung der Abstände der jeweiligen Punkte zu ihrem Schwerpunkt:

$$s = \frac{\sum_i \|\tilde{a}_i\| \|\tilde{b}_i\|}{\sum_i \|\tilde{a}_i\|^2} \tag{4.31}$$

Die Punkte $\tilde{b}_i$ werden in eine Matrix $\boldsymbol{B}$ und die skalierten Punkte $s\tilde{a}_i$ in eine Matrix $\boldsymbol{A}$ spaltenweise zusammengefügt. Die gesuchte Rotationsmatrix muss

$$\tilde{b}_i = s\boldsymbol{R}\tilde{a}_i \forall i \tag{4.32}$$

erfüllen. 4.32 ist bekannt als das *Orthogonal Procrust Problem*. Die optimale Rotation zwischen den Punktemengen $\boldsymbol{B}$ und $\boldsymbol{A}$ ist gegeben durch:

$$\boldsymbol{R} = \boldsymbol{V}_R \boldsymbol{U}_R{}^{\mathrm{T}} \tag{4.33}$$

$\boldsymbol{V}_R$ und $\boldsymbol{U}_R$ sind die rechten und linken Singulärvektoren der Singulärwertzerlegung von $\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}}$. Diese Rotationsmatrix ist per Definition nun orthonormal. Im Falle schlechter Datenmengen, die extrem verrauscht sind, wird die so bestimmte Matrix eher orthonormal sein als die Rotationsbeziehung zwischen den Punktmengen zu beschreiben.

Als einen letzten Schritt gilt es die optimale Translation zu ermitteln durch:

$$\boldsymbol{t} = s^{-1}\boldsymbol{R}^{\mathrm{T}} b_0 - a_0 \tag{4.34}$$

Die finale Pose ist gegeben durch die orthonormale Rotation und die darauf basierende Translation:

$$\boldsymbol{P} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \tag{4.35}$$

4.2.7 Kameraposition und Projektionsmatrix

Der Zusammenhang zwischen Lage des optischen Zentrums einer Kamera und der Kamerapose ist folgend beschrieben.
Ein Punkt $\tilde{\boldsymbol{X}}_w^{\mathrm{w}}$ im Weltkoordinatensystem ist durch die Transformation

$$\tilde{\boldsymbol{X}}_c^{\mathrm{w}} = \boldsymbol{R}\tilde{\boldsymbol{X}}_w^{\mathrm{w}} + \boldsymbol{t} \tag{4.36}$$

in das jeweilige Kamerakoordinatensystem zu überführen.
Aus 4.36 lässt sich das optische Zentrum C der Kamera errechnen als:

$$0 = \boldsymbol{R}C + \boldsymbol{t} \tag{4.37}$$

und

$$C = \boldsymbol{R}^{-1} - \boldsymbol{t} = \boldsymbol{R}^{\mathrm{T}} - \boldsymbol{t} \tag{4.38}$$

Die Lage des Kamerazentrums im globalen Koordinatensystem entspricht der Rotation des Referenzkoordinatensystems mit der inversen Rotationsmatrix gefolgt von der negativen Translation der Projektionsmatrix.

Kapitel 5

Experimente und Ergebnisse

In diesem Kapitel sind die Experimente und Ergebnisse, die während der Arbeit entstanden sind, aufgezeigt. Sowohl die Kameramatrizen der Modellbilder als auch die jeweils angegebenen Posen wurden mit dem in 4.2.6 beschriebenen Verfahren bestimmt.
In 5.1 werden die generierten Modelle von realen Szenen präsentiert. Dabei ist je eine Poseschätzung einer Kameraposition, die nicht aus der Modellserie stammt, exemplarisch mit angegeben.
In 5.2 werden die Ergebnisse anhand eines synthetischen Modells mit Ground Truth präsentiert.
In 5.3 werden einige Fälle aufgezeigt, die zu Schwierigkeiten während der Modellerstellung und der Posebestimmung führen.

5.1 Reale Szenen

Sämtliche Bilder der hier aufgeführten Szenen und Posen wurden mit einer digitalen Spiegelreflexkamera Nikon D60 bei einer Brennweite von $18mm$ erstellt. Die Auflösung der Bilder beträgt horizontal 3872 und vertikal 2592 Pixel. Die radialen Verzerrungseffekte der Kamera wurden im Vorfeld des Programmablaufs ausgeglichen. Die Bildserien der Modelle sind freihändig und ohne Maß in der Rotation oder Translation zwischen den Aufnahmen angefertigt.
Ein Überblick über die Anzahl der Modellbilder, die Anzahl der triangulierten Weltpunkte und die Menge der SIFT Merkmale der realen Modelle ist in Tabelle 5.1 zusammengefasst. Die Tabelle gibt dabei auch die durchschnittliche Anzahl an SIFT-Features pro Weltpunkt wieder (FpW). Diese Zahl ist ein Indiz für den Beschreibungsradius der Weltpunkte. In ihren Spitzenwerten wurden in der Szene *Deutsches Eck* sogar einige Weltpunkte aus allen 31 Modellbildern beschrieben.

Szene	Frames	Weltpunkte	Features	ØFpW
Deutsches Eck	31	14243	41392	2,91
Sparkassenhaus	18	8590	21900	2,55
Pforte	14	3201	7464	2,33
Kaiserin Augusta	43	16730	44015	2,63

Tabelle 5.1: Szenen in Zahlen

Die Abbildungen in den Kapiteln 5.1.1 bis 5.1.4 sind wie folgt zu verstehen:

a: zeigt jeweils das Modell in der *OpenGL* Ansicht. Die magentafarbenen Kameraframes entsprechen dabei den Kameraposen der Bildserie, die zur Modellerstellung benutzt wurden. Der grüne Kameraframe ist die Pose einer Aufnahme, die nicht aus der Modellserie stammt. Sie ist die eigentlich gesuchte Pose dieser Arbeit. Neben den weiß dargestellten Weltpunkten werden die 3D-Punkte, die zur Poseschätzung der grün markierten Kamera verwendet werden ebenfalls grün hervorgehoben.

b: zeigt das Bild, dessen Pose gesucht ist. Neben den roten SIFT-Merkmalen sind die zu den 3D-Positionen im Modell gehörigen 2D-Bildpositionen grün gefärbt. Die halbtransparente blaue Umrandung dieser Bildpositionen entspricht dem Rückprojektionsfehler der dreidimensionalen Punkte mit der aktuell geschätzten Projektionsmatrix in das Posebild.

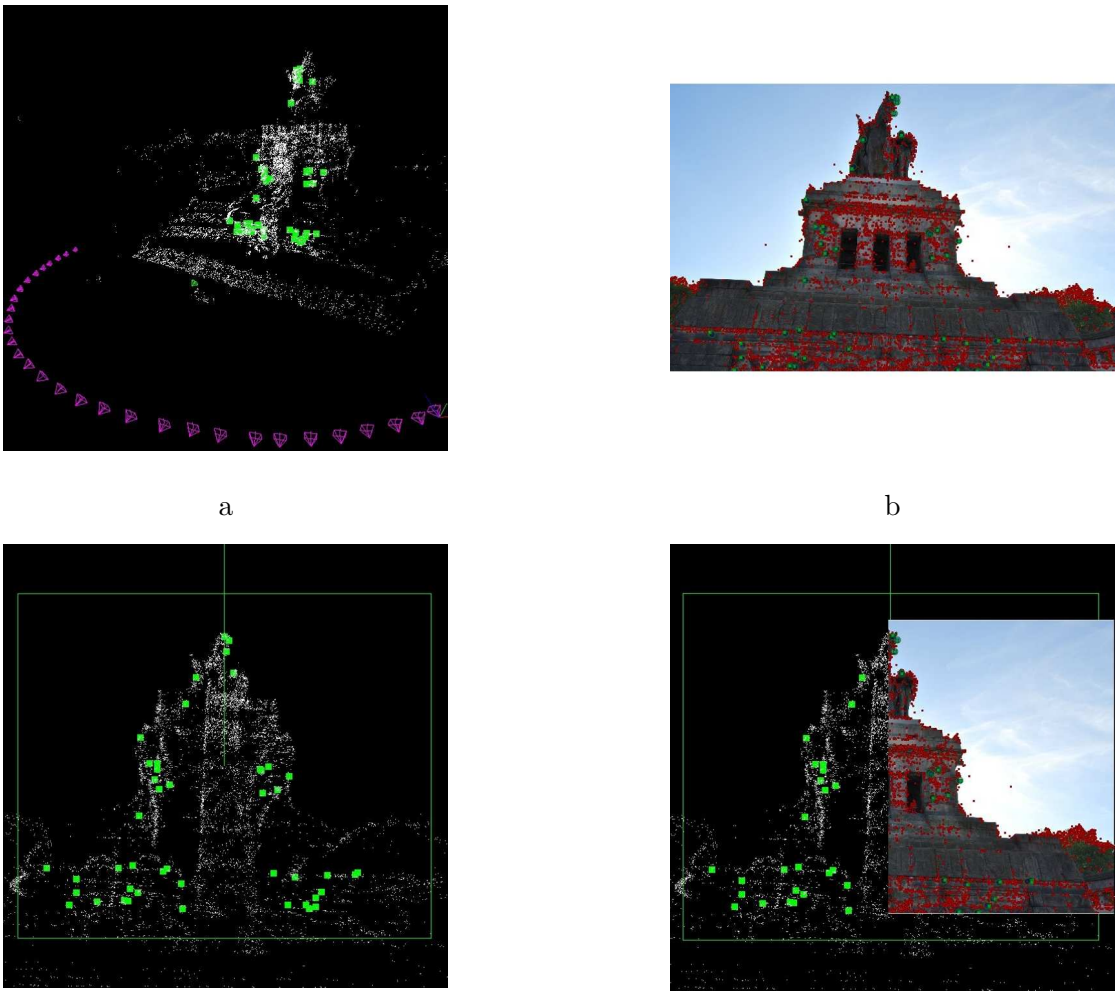
c: stellt die Ansicht aus der errechneten Pose im dreidimensionalen Modell dar.

d: visualisiert eine Überlagerung des Posebildes und des Modells aus der Poseperspektive.

5.1.1 Deutsches Eck

Die Szene *Deutsches Eck* (siehe Abbildung 5.1) besteht aus 31 Einzelaufnahmen, die jeweils eine relative Blickwinkeländerung von etwa 5° aufweisen. Eine Drehung von circa 180° um das rigide Objekt ist somit in dem Modell beschrieben. In Abbildung C.1 ist ein Ausschnitt von 2 Bildern, die zur Erstellung des Modells dienen, gezeigt. Zum Zeitpunkt der Modellerstellung herrschte eine nicht unerhebliche partielle Verdeckung der Szene in Form eines Baugerüsts vor. Die Aufnahmen zur Poseschätzung sind ohne diese Verdeckung entstanden. Außerdem ist eine sehr unterschiedliche Beleuchtungssituation zwischen der Modellserie und den Posebildern gegeben.

Durch die dichte Abtastung der Szene und die Menge an SIFT Features (es sind insgesamt 41392 Merkmale an 14243 Weltpunkte annotiert) ist es trotz der partiellen Verdeckung möglich, stabil Posebilder gegen das Modell zu vergleichen. Die natürlichen Beleuchtungsänderungen werden durch die Eigenschaften des SIFT Deskriptors (siehe 3.1.5) ebenfalls gut ausgeglichen.



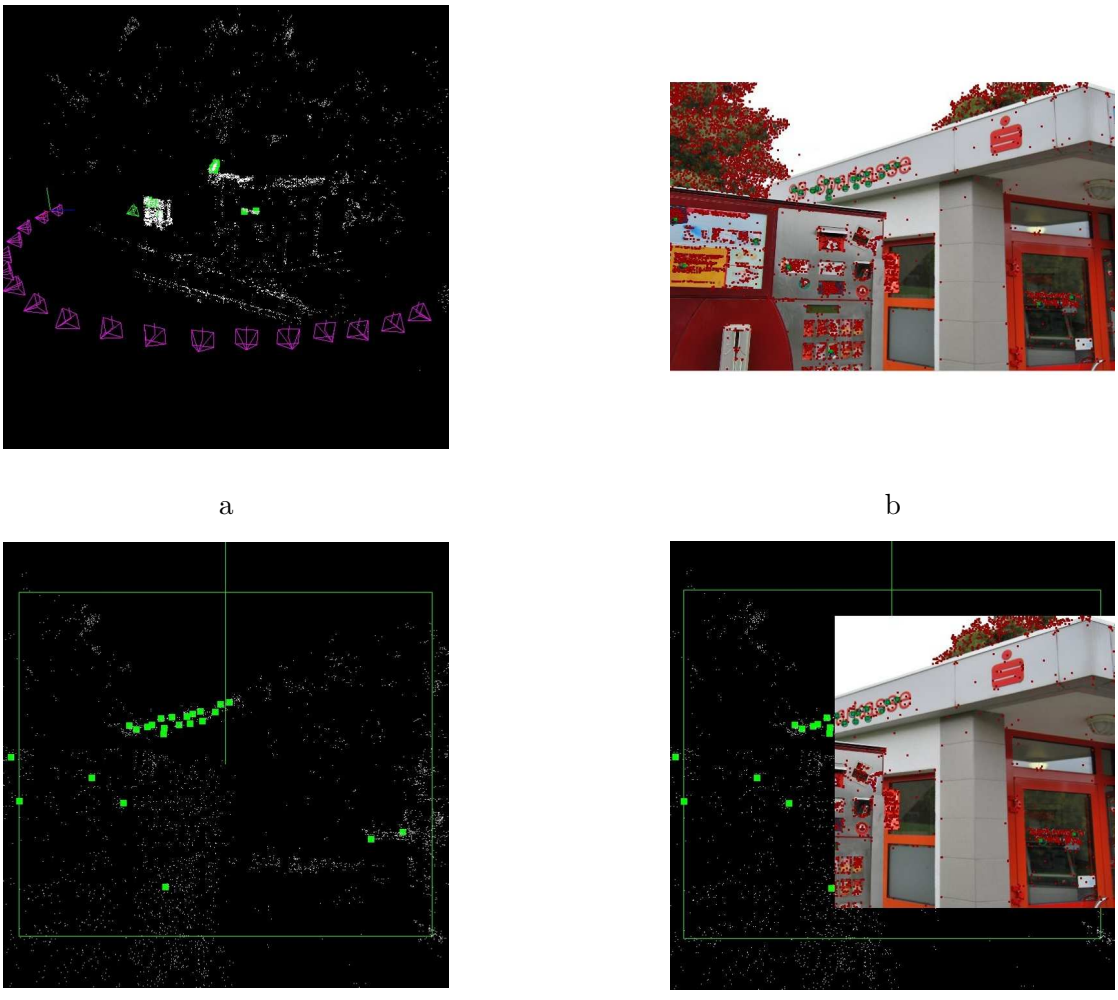
a b

c d

Abbildung 5.1: Szene 'Deutsches Eck'

5.1.2 Sparkassenhaus

Die Szene *Sparkassenhaus* (siehe Abbildung 5.2) setzt sich aus 18 Einzelbildern zusammen. Mit einer relativen Blickwinkeländerung von etwa 10° ergibt sich ein Beschreibungsradius von ungefähr 180° der betrachteten Geometrie. Das Posebild zeigt eine relativ nahe an der Struktur des Modells lokalisierte Kameraposition. Sie wird darüber hinaus von dominanten Ebenen im Bild geprägt auf denen sich viele Posemerkmale befinden. Die meisten der verwendeten Posemerkmale sind aus dem Schriftzug der linken Fassade extrahiert. Trotz dieser Umstände ist die Poseschätzung anhand von 3D-2D-Korrespondenzen stabil erfolgt.



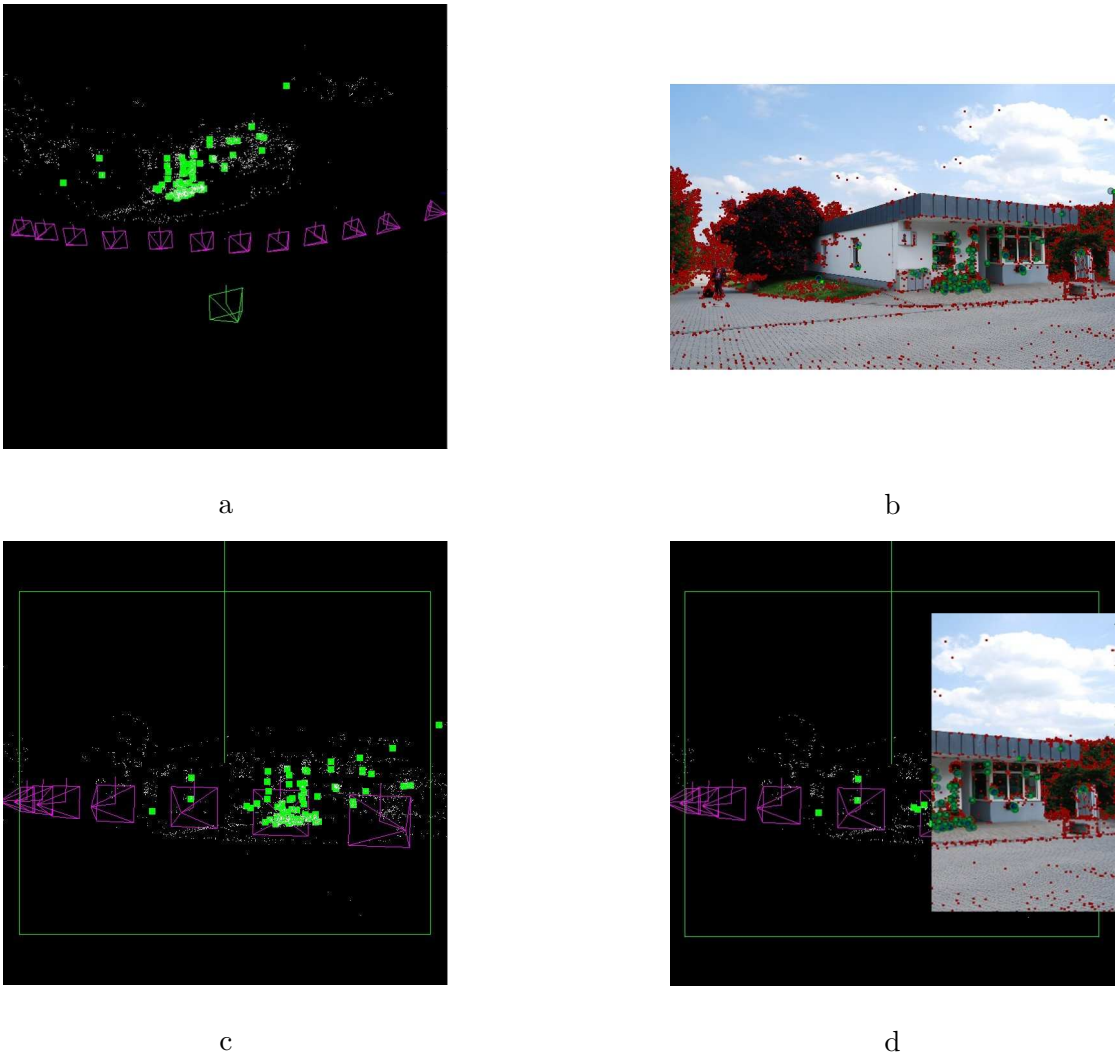
a b

c d

Abbildung 5.2: Szene 'Sparkassenhaus'

5.1.3 Pförtnerhaus

Die Szene *Pförtnerhaus* (siehe Abbildung 5.3) besteht aus 14 Einzelbildern aus denen 7464 Bildmerkmale 3201 Weltpunkten zugeordnet sind. Die Aufnahmen repräsentieren eine Bewegung von ungefähr 90° um eine Ecke der vorderen Gebäudefronten. Die Merkmale, die zur Schätzung der Pose herangezogen wurden liegen auf beiden Gebäudefronten verteilt. Der Hauptteil konzentriert sich aber auf der vorderen Front, die durch ihre Texturierung wesentlich dichter im Modell beschrieben ist. Trotz der stark lokalen Konzentration der Posemerkmale ist eine gültige Schätzung der externen Orientierung erfolgt.



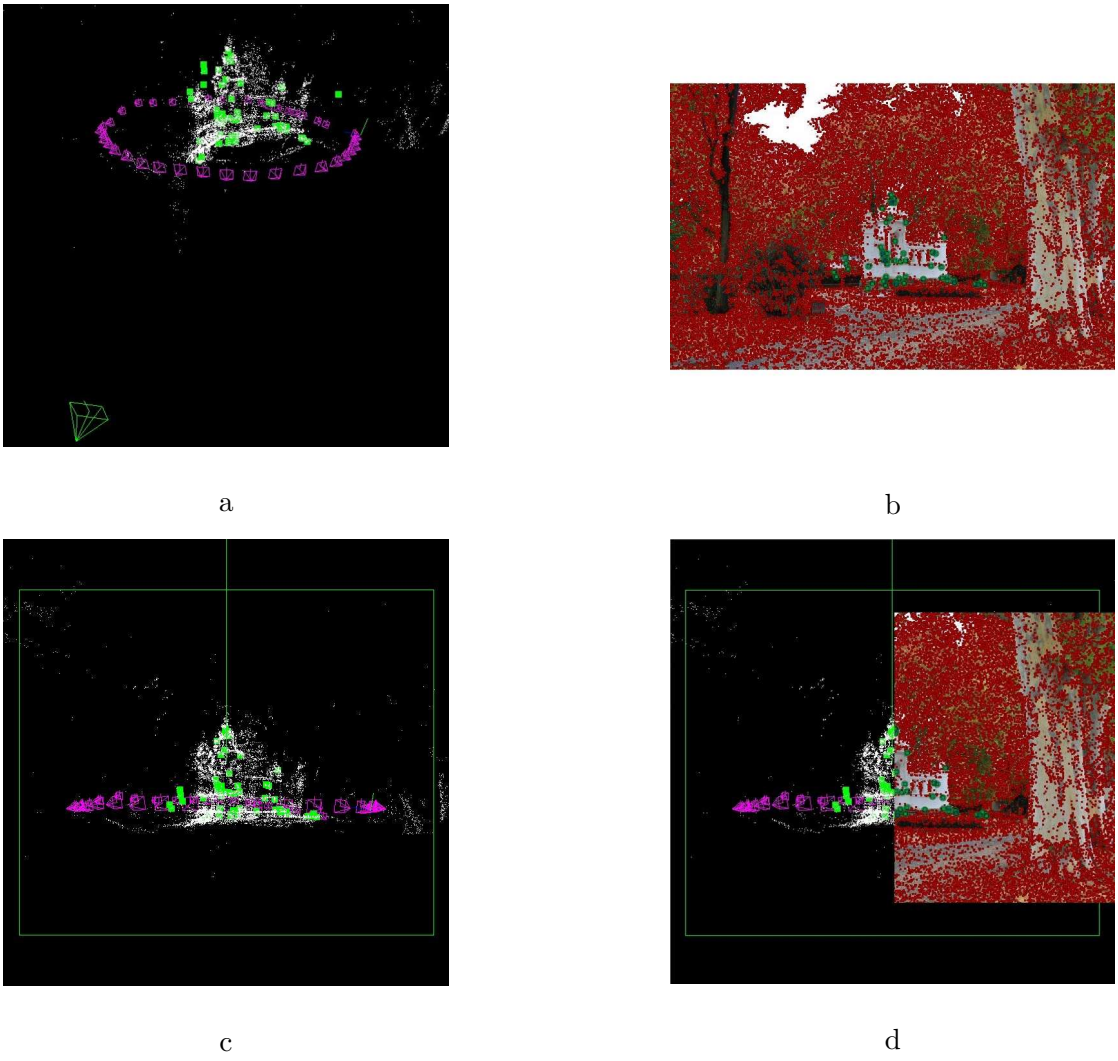
a b

c d

Abbildung 5.3: Szene 'Pforte'

5.1.4 Kaiserin Augusta

Die Szene *Kaiserin Augusta* (siehe Abbildung 5.4) stellt eine komplette 360° Ansicht dar. Sie besteht aus 43 konsekutiven Aufnahmen, was einer durchschnittlichen Blickwinkeländerung von etwa 8° entspricht. Die geschätzte Kamerapose weist eine beträchtliche Entfernung zum Modell auf und ist dennoch stabil wiedererkannt worden. Das Posebild hat insgesamt 42651 Posemerkmale. Ein Großteil derer entsteht aus sehr selbstähnlichen Merkmalen in Form von Laub und wird bereits anhand des Distance-Ratio-Kriteriums verworfen. Die restlichen Posemerkmale sind stabil genug, eine robuste Schätzung auf diese Distanz zu ermöglichen.



a b

c d

Abbildung 5.4: Szene 'Kaiserin Augusta'

5.2 Ground Truth Test Campusmodell

Die Korrektheit der real erstellten Modelle und deren Pose ist nur visuell zu überprüfen. Da die Aufnahmen frei Hand und ohne ein Maß erstellt wurden, lässt sich die aktuelle Projektionsmatrix nicht metrisch verifizieren.
In einem Projektpraktikum *Markante Merkmale II* der Universität Koblenz unter Leitung von Prof. Dr. Lutz Priese und Frank Schmitt entstand ein Modell des Campus Koblenz, das metrisch korrekt und äußerst detailreich ist. Es eignet sich deshalb zu einer *Ground-Truth-Überprüfung* der hier entstandenen Vorgehensweise. Das *Campusmodell* wurde mit der 3D-Grafik-Software *Blender* erstellt. Aus diesem Modell wurden mit vorgegebener Position und Orientierung eine Serie von neun Bildern gerendert (siehe Abbildung 5.5). Diese Bildserie ist die Basis für die Modellerstellung.

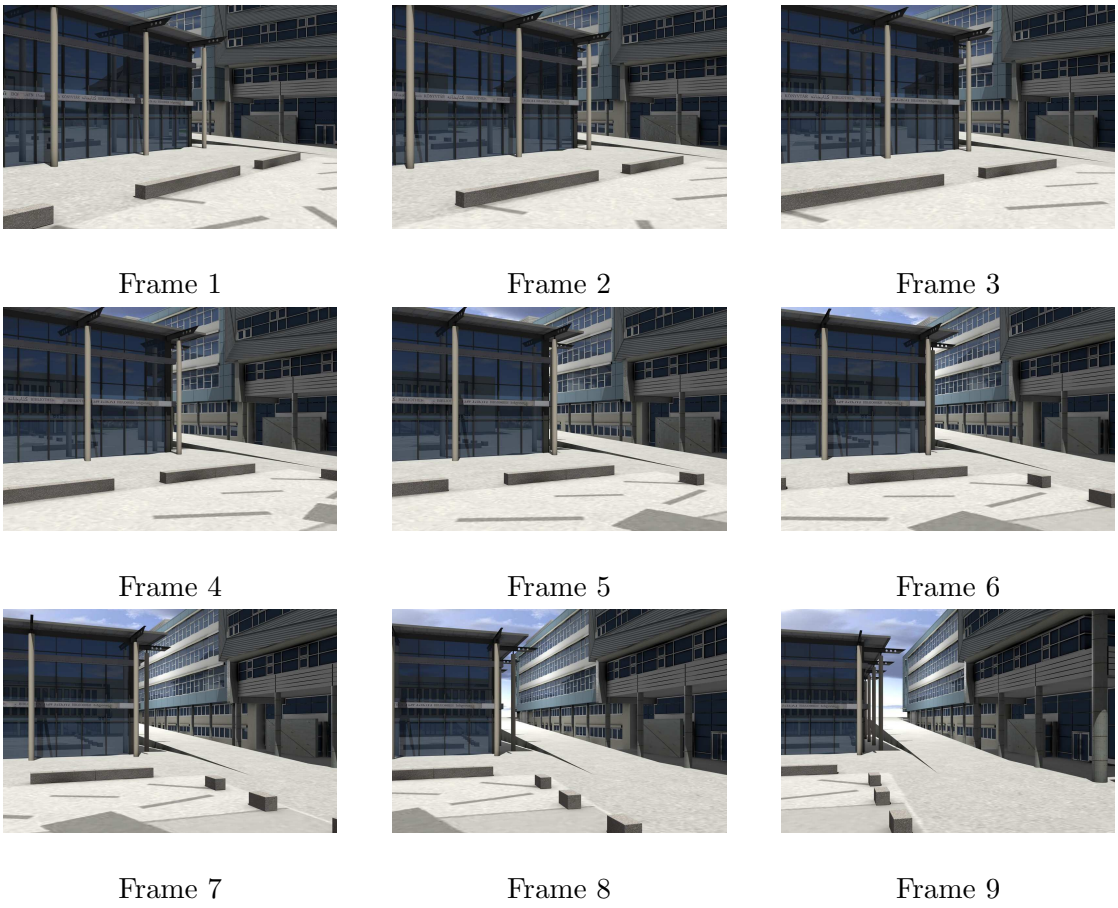


Abbildung 5.5: Bildserie 'Campusmodell'

Frame	Referenz $(\boldsymbol{R}_x, \boldsymbol{R}_y, \boldsymbol{R}_z)$	$\Delta\boldsymbol{R}_x$	$\Delta\boldsymbol{R}_y$	$\Delta\boldsymbol{R}_z$
1	0, 0, 0	0	0	0
2	0, 5, 0	-0.0019431	0.26690	0.023798
3	0, 10, 0	0.010205	0.58860	0.010523
4	0, 15, 0	-0.023295	0.99400	-0.002691
5	0, 20, 0	-0.045454	1.4250	0.037656
6	0, 25, 0	-0.030474	1.9320	0.094671
7	0, 30, 0	-0.0071658	2.5070	0.15866
8	0, 32.5, 0	-0.15198	2.9290	0.070525
9	0, 35, 0	-0.047676	3.3200	0.20806

Tabelle 5.2: Modellabweichungen in der Rotation

Das Resultat der Modellerstellung ist in Abbildung 5.7 dargestellt. In der oberen der beiden Ansichten ist das errechnete Modell zu sehen, die untere Ansicht zeigt die Positionen der Kameras aus der Grafik-Software *Blender*. In dem Modell sind 1577 Weltpunkte beschrieben mit einer globalen Keylistengröße von 4100 Merkmalen.

Die intrinsischen Parameter der *Blender-Kamera* werden als ideal angenommen. Bei einer Auflösung von 3872×2592 ist die Kalibrier-Matrix mit einem 1:1 Verhältnis der Pixelbreite und Pixelhöhe ($\alpha = \beta$), einem zur Bildachse parallel ausgerichtetem Bildsensor ($\lambda = 0$) und dem Kamerahauptpunkt im Zentrum des Bildes gegeben als:

$$\boldsymbol{K} = \begin{bmatrix} 3872.0000000000 & 0.0000000000 & 1936.0000000000 \\ 0.0000000000 & -3872.0000000000 & 1296.0000000000 \\ 0.0000000000 & 0.0000000000 & 1.0000000000 \end{bmatrix} \tag{5.1}$$

Eine Poseschätzung kann nur so korrekt wie das Modell sein, auf dem sie basiert. Deshalb werden zunächst die Projektionsmatrizen der einzelnen Modellbilder verifiziert. Tabelle 5.2 zeigt die Abweichungen der Rotationsmatrizen aus dem Modell von ihrem Idealwert in Grad. Die Winkel stellen die Euler-Winkel der entsprechenden Drehmatrix dar. Die Kameras wurden dabei um ihre Hochachse rotiert und in der Ebene transliert. Der Idealwert ist bestimmt durch die Rotations- und Translationsparameter, die beim Rendern der Bilder angegeben wurden. Das Koordinatensystem liegt als Rechtssystem vor mit einer horizontalen X-Achse, einer vertikalen Y-Achse und der Blickrichtung entlang der negativen Z-Achse.

Mit zunehmender Entfernung der Kameras von ihrem Weltursprung ist eine zunehmende Abweichung in der Rotation festzustellen. Ein Maximalwert von $3,3200°$ über die gesamte Rekonstruktionsfolge wurde ermittelt.
Die relative Translation der Kameras folgt auf deren Rotation. Der begangene Fehler ist deshalb direkt abhängig von der Abweichung in der Rotation. Er ist jedoch wesentlich geringer als die Rotationsunterschiede. Für Frame 6 ist die ideale Tranlsation in Blickrichtung 0. Das bedeutet, dass die Position der Kamera genau auf der um 25° um die Y-Achse der ersten Kamera rotierten Geraden entlang der so gedrehten X-Achse liegt. Eine Distanz von $0,0648871$ in Blickrichtung von der tatsächlichen Position wurde ermittelt. Die Abweichungen in der Y-Achse sämtlicher Kameras betrugen einen maximalen Wert von $0,015226$. Die relativen Translationen in X-Richtung sind sowohl in ihrem Idealwert als auch in den Modellwerten äquidistant. Auf die gesamte Rekonstruktionsserie hat sich eine Abweichung von $0,13359$ in der relativen X-Richtung des Frames 9 akkumuliert.
Zur Überprüfung der Pose einer Kamera, die nicht aus der Bildserie stammt, wurden zwei Bilder der in dem Modell beschriebenen Szene gerendert (siehe Abbildung 5.6).

Pose1

Pose2

Abbildung 5.6: Posen des Campusmodells

Wie in 4.2.3 erwähnt basiert die Schätzung der Pose auf dem Support Set eines RANSAC-Durchlaufs. Das bedeutet, dass die Poseschätzung je nach zurückgelieferten Daten einen leichten Versatz beinhaltet. Aus diesem Grund wurden aus je vier Poseschätzungen die Mittelwerte der Rotationswinkelabweichungen aufgerechnet. Die höchste Varianz der Werte ist in der X-Rotation etwa $0,6°$, in der Y-Rotation 1° und in der Z-Rotation $1,2°$. Das Resultat für beide Posebilder zeigt Tabelle 5.3.

Die Ergebnisse der Poseschätzung decken sich mit den Ergebnissen der Modellerstellung. Die angegebenen Abweichungen sind also im Modell begründet.

Frame	Referenz $(\boldsymbol{R}_x, \boldsymbol{R}_y, \boldsymbol{R}_z)$	$\Delta\boldsymbol{R}_x$	$\Delta\boldsymbol{R}_y$	$\Delta\boldsymbol{R}_z$
Pose1	0, 15, 0	0.54214	2.6220	0.24886
Pose2	0, -5, 0	0.39137	2.8160	0.73246

Tabelle 5.3: Mittelwerte der Poseabweichung

Die entsprechenden Translationen sind analog zu den Ergebnissen der Modellerstellung auf den Rotationsergebnissen basierend. Sie weisen bei der Poseschätzung eine sehr geringe Varianz auf und sind daher als stabil zu betrachten.
Insgesamt ist das *Campusmodell* wegen seinen stark dominanten Flächen der Gebäudefronten und den relativ wenigen Features, die stabil wiedergefunden werden, ein anspruchsvolles Szenario. Die ermittelten Werte sollten deshalb als äußerste Grenzen des Verfahrens angesehen werden. In den präsentierten realen Szenen sind Modellaufbau und Poseschätzung als korrekter anzusehen.

5.3 Problemfälle des Verfahrens

In diesem Kapitel werden einige Problemfälle aufgezeigt.
Während der Modellerstellung ist darauf zu achten, dass in den fortlaufenden Bildern genügend Anschlussfeatures vorhanden sind. Sind die Regionen, die sich in beiden Bildern überschneiden zu klein oder existieren zu wenige Features durch eine homogene Texturierung, so ist der sequentielle Ablauf der Modellerstellung unterbrochen. Extrem dominante Flächen, die selbst per manueller Auswahl keine gute Basis für den 8-Punkte-Algorithmus liefern, stellen ebenfalls ein Problem dar. Zu lange Bildserien erweisen sich noch immer als schwierig. Trotz der deutlichen Verbesserung gegenüber der E-Matrix-Zerlegung haben sich Fehler im Modell gebildet. Diese sind zwar relativ gering, aber dennoch störend.
Die Grenzen der Poseschätzung sind direkt mit den Eigenschaften der SIFT-Features verknüpft. [Low04] zeigt, dass bei einer Blickwinkelneigung zu einer planaren Fläche von 30° circa 80% der Merkmale stabil verglichen werden können. Bei einer Neigung von 50° fällt dieser Wert schon auf unter 50%. In der Praxis bedeutet dies, dass je nach Menge der zur Verfügung stehenden Merkmale eine Blickwinkelneigung von etwa 30° über die äußeren Grenzen des Modells nicht überschritten werden sollte.
Ein generelles Problem entsteht aus zu wenig Korrespondenzen zwischen Modell und Posebild. Die Anzahl der minimalen Modellparameter muss stets gewährleistet sein. In den hier verwendeten linearen Optimierungsverfahren werden minde-

stens 6 Welt-Bild-Zuordnungen erwartet. Sie müssen korrekt sein, um eine gültige Posebestimmung durchzuführen. Praktisch sind viele der Punkte ungeeignet und der RANSAC-Algorithmus benötigt deutlich mehr Korrespondenzen, um ein gutes Support Set zu finden.

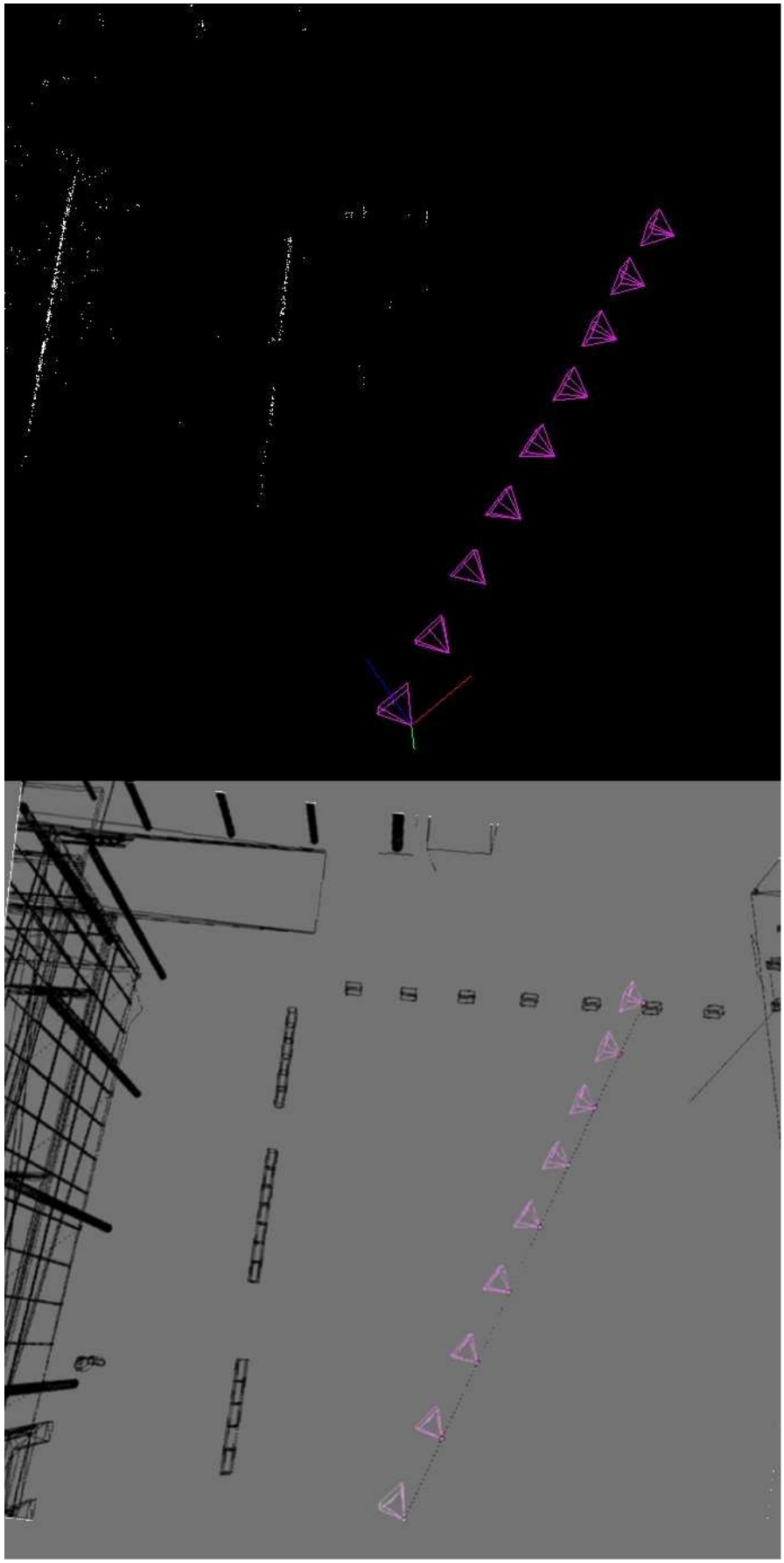


Abbildung 5.7: Oben: Modell in der *OpenGL* Visualisierung, Unten: Renderansicht aus *Blender*

Kapitel 6

Zusammenfassung und Ausblick

In diesem Kapitel wird der Ablauf der Arbeit und die wichtigsten Erkenntnisse, die daraus resultieren zusammengefasst. Ein Ausblick auf mögliche Erweiterungen folgt anschließend in 6.2.

6.1 Zusammenfassung

In der Arbeit *Modellbasierte Posebestimmung aus 2-D/3-D SIFT-Korrespondenzen* entstand ein Verfahren, das die Pose eines Bildes anhand der Korrespondenzen zwischen einem zuvor erstelltem Modell der Szene und den aus dem Posebild extrahierten Bildmerkmalen bestimmt.
Es wurde zunächst die Motivation für die Wahl der Bildmerkmale dargelegt und die Eigenschaften der Scale Invariant Feature Transform hervorgehoben. Sämtliche Bilder werden mit einer handelsüblichen monokularen Digitalkamera erstellt, deren intrinsische Parameter in einem Kalibrierschritt zuvor ermittelt wurden.
Ist eine Aufnahmeserie einer rigiden Szene frei Hand erstellt, wird daraus das Modell generiert. Der erste Schritt dabei ist die Extraktion der SIFT-Merkmale der Eingabebilder. Die automatische Suche eines nächsten Nachbarn zwischen den Bildmerkmalen der konsekutiven Aufnahmen liefert die Grundlage zur Lösung des Korrespondenzproblems. Der Aufwand der linearen Suche wurde durch die Integration einer Suchmethode, die auf Partitionierung des Merkmalsraumes beruht, erheblich minimiert. Ein Distance-Ratio-Kriterium, das ungeeignete Merkmale schon im Vorfeld der Anwendung eliminiert, ist die erste Maßnahme zur Steigerung der Robustheit der Korrespondenzbestimmung.
Die verbleibenden Punkte werden anhand der epipolaren Bedingung zwischen deren Ursprungsbildern auf Ausreißer überprüft. Dies geschieht mittels einer manuell geschätzten Fundamental-Matrix. Das Problem der degenerierten Konfigurationen zur automatischen Schätzung der Fundamental-Matrix wurde dargelegt und der

gewählte manuelle Ansatz begründet.
Aus der Fundamental-Matrix wird die Essential-Matrix extrahiert, die zur Initialisierung der Weltpunkte des Modells dient. Somit ist der Grundstein des Modells gelegt und es können beliebig viele Anschlussbilder angefügt werden.
Es wurde zunächst versucht, über eine Zerlegung der E-Matrizen zwischen weiterführenden Frames das Modell zu konstruieren. Dabei haben sich jedoch Unstimmigkeiten in dem Translationsbetrag der Kameras zueinander als hinderlich erwiesen. Fehler, die sich ergeben werden durch den rekursiven Aufbau in alle weiteren Kameramatrizen mit einfließen und sich bei weiteren Fehlschätzungen verstärken.
Die Symmetrie- und Transitivitätseigenschaften der Korrespondenzrelation ermöglichen es, direkt die globale Projektionsmatrix eines anschließenden Bildes aus den bisherigen dreidimensionalen Punkten und ihren Zuordnungen im Bild zu errechnen. Das Problem der inkonsistenten Kameramatrizen wurde gelöst und die Modelle so erheblich verbessert. Das Verfahren wird analog bei der folgenden Posebestimmung angewandt.
Besteht bereits ein korrespondierendes Merkmal in vorangegangenen Frames, werden die zusätzlichen Informationen aus den neuen Bildern an den existierenden Weltpunkt angetragen. Die neuen Korrespondenzen werden in Weltpunkte trianguliert und ebenfalls in das Modell aufgenommen. Somit ist gewährleistet, dass ein Weltpunkt auch als physische Entität behandelt wird, die aus mehreren Ansichten beschrieben ist. Dies steigert den Betrachtungsradius, aus dem ein Weltpunkt während der Poseschätzung erscheint, erheblich. Es wurden so Merkmale festgestellt, die sich über eine beachtliche Serie von Bildern wiederfinden lassen. In der in 5.1.1 gezeigten Szene wurden einzelne Weltpunkte durch Deskriptoren aus allen Modellbildern beschrieben.
Auf diese Art wurden Modelle realer Szenen von bis zu 360° aus 43 Einzelaufnahmen erstellt. Alle Informationen, die nach der Modellgenerierung zur Verfügung stehen wurden aufgezeigt und das Modell formal definiert.
Die Poseschätzung erfolgt in einem zweiten Schritt. Über eine Zuordnung der Merkmale des Posebildes und der Merkmale des Modells wird die dreidimensionale Position eines Bildmerkmals erfasst. Über diese 3D-2D-Informationen wird die externe Orientierung der Kamera zum Zeitpunkt der Poseaufnahme errechnet. Die automatischen Zuordnungen der Posemerkmale und der Modellmerkmale werden dabei anhand des bekannten Distanz-Verhältnis-Kriteriums überprüft. Punkte, die mehrfach auf einer lokalen Bild- und Weltposition liegen erwiesen sich als ungeeignet. Sie werden in einem folgenden Schritt eliminiert.
Die automatischen Zuordnungen sind jedoch fehlerbehaftet. Ein RANSAC-Verfahren ermittelt deshalb die Inlier, die zur Schätzung der neuen Projektionsmatrix herangezogen werden.

Zwei lineare Verfahren wurden zur Poseschätzung angewendet. Die direkte lineare Transformation erwies sich als sehr instabil, da sie die Orthonormalitätsbedingung einer Rotationsmatrix vernachlässigt und alle Poseparameter unabhängig voneinander schätzt. Das Ergebnis kann nicht ohne Weiteres als eine gültige Pose angenommen werden. Es wurde daraufhin die nächste Rotationsmatrix einer DLT-geschätzten Rotation bestimmt. Durch die nachträgliche Korrektur einer Rotation verändert sich aber folglich auch die Lage der Kamera im Raum und es entstehen Rückprojektionsfehler. Sie könnten eventuell iterativ ausgeglichen werden. Ein iterativer Ansatz wird aber hier wegen des Aufwands als keine passende Lösung betrachtet.
Ein zweites lineares Verfahren nach [Fio01], das die Kamerapose in einer geschlossenen Form berechnet, wurde deshalb integriert. Dabei werden zunächst die tatsächlichen Tiefenwerte der Weltpunkte aus Kamerasicht und die Skalierung des Modells bestimmt. Daraufhin lässt sich eine orthonormale Rotation in die Punktemengen einpassen und die Translation zwischen den Kameras als ein letzter Schritt errechnen. Dieses Verfahren erwies sich als wesentlich stabiler, da die Rotation und Translation der Kameramatrix aus den Schwerpunkten der Bild- und Weltpunktmengen bestimmt wird.
Die entstandenen Modelle und einige darauf basierende Poseschätzungen wurden präsentiert und das Ergebnis anhand eines synthetischen Modells auf dessen Ground Truth getestet. Das Fazit der Modellerstellung ist dabei, dass sich noch kleine Anschlussfehler in langen Bildserien einschleichen können, die einen eventuellen Anlass zu Optimierung geben (siehe 6.2). Die Modelle waren jedoch dicht repräsentiert und visuell stimmig mit der dargestellten Szene. Die Poseschätzung hat sich insgesamt als stabil herausgestellt. Es wurden sowohl nahe an der Szene lokalisierte als auch weit davon entfernte Kameraposen zufriedenstellend geschätzt. Die Grenzfälle des Ansatzes wurden dargelegt und begründet. Das abschließende nächste Kapitel gibt einige Vorschläge zur Optimierung des bestehenden Verfahrens.

6.2 Ausblick

Der bisherige Ablauf der Modellerstellung ist stark sequentiell ausgerichtet. Kommt es zu Anschlussproblemen zwischen fortlaufenden Frames, so versagt das Verfahren. Eine Möglichkeit, hier flexibler zu reagieren würde die Modellerstellung wesentlich optimieren. [SZ02] zeigt eine mögliche Lösung dieses Problems. Die Eingabebilder werden in einem Spanning Tree nach der Anzahl ihrer Korrespondenzen strukturiert. Während der Modellerstellung ist so das optimale nächste Bild zu bestimmen.
Ein Ansatz zur Steigerung der Korrektheit der Modelle wäre in einem Bundle-

Adjustment-Verfahren zu sehen. Hier müsste überprüft werden, ob eine globale Optimierung nach der Modellerstellung ausreichend ist oder jede neu eingefügte Kameramatrix optimiert werden sollte. Die Basis für beide Herangehensweisen ist gegeben.
In wie weit eine solch dichte Repräsentation der dargestellten Szenen notwendig ist, liegt sicherlich an der Anwendung selbst. Für eine Poseschätzung sind wesentlich weniger Punkte ausreichend als sie im Moment verwendet werden. Eine Ausdünnung des Modells wäre anzudenken.
Die Orientierung und Skala der SIFT-Merkmale bietet weitere Optimierungsansätze. Eine geschätzte Pose muss insgesamt stimmig mit den Unterschieden in Orientierung und Skala der Modellmerkmale und der Posemerkmale sein. Zu große Abweichungen zwischen ihnen würden auf eine falsch geschätzte Pose hinweisen.
Ein Indiz für eine falsche Pose wäre außerdem in der Tatsache zu sehen, dass ein zur Poseschätzung herangezogener Weltpunkt zwingend im Sichtfeld der zu schätzenden Kamera liegen muss. Ist dies durch die Lage der Kamera ausgeschlossen, könnten die Punkte verworfen und Grenzfälle stabiler gestaltet werden.
Eine mögliche Steigerung der Stabilität ließe sich auch dadurch gewährleisten, dass nur Weltpunkte für die Poseschätzung verwendet werden, die aus vielen Modellansichten beschrieben sind. Solche Punkte gelten als sehr zuverlässig. Die vorgestellte Modellstruktur biete die dafür benötigte Funktionalität.
Verfahren, die weniger als 6 Korrespondenzen benötigen (zum Beispiel [FB81] und [QL99]) würden die Anzahl der benötigten RANSAC-Iterationen erheblich minimieren und somit in einer kürzeren Laufzeit resultieren.
Als ein letzter Optimierungsansatz wären geeignete Mittel zu verwenden, die eine manuelle Selektion der Merkmale zur Bestimmung der Fundamental-Matrix ersetzen. Eine automatisierte Modellerstellung wäre somit denkbar.

Anhang A

Symbole und Bezeichner

In nachfolgender Tabelle sind die in der vorliegenden Arbeit verwendeten Symbole und Bezeichner sowie ihre jeweilige Bedeutung aufgelistet.

Symbol/Bezeichner	Bedeutung
$\tilde{\boldsymbol{p}}^{\mathrm{i}}$	Ein Punkt in homogenen Bildkoordinaten
$\tilde{\boldsymbol{p}}^{\mathrm{w}}$	Ein Punkt in homogenen Weltkoordinaten
$\tilde{\boldsymbol{p}}^{\mathrm{p}}$	Ein Punkt in homogenen Pixelkoordinaten
θ	Der Schwellwert für die Epipolare Bedingung
Θ	Die Orientierung eines SIFT-Features
$\boldsymbol{E}$	Eine Essential-Matrix
$\boldsymbol{F}$	Eine Fundamental-Matrix
$\boldsymbol{K}$	Eine Kalibrier-Matrix
α, β	Die Skalierungsfaktoren der Pixelbreite und Pixelhöhe in u und v Bildkoordinaten
λ	Der Winkel zwischen dem Bildsensor und dem Kamerakoordinatensystem
$\boldsymbol{I}$	Die Identitäts-Matrix
$\boldsymbol{P}$	Eine Projektionsmatrix
$\boldsymbol{R}$	Eine Rotationsmatrix
$\boldsymbol{t}$	Ein Translationsvektor
s	Der Skalierungsfaktor eines Modells
F_i	Der Frame der Kamera i
r	Der Schwellwert der Abstandsverhältnisse zweier nächster Nachbarn

Fortgesetzt auf der nächsten Seite

Fortsetzung der vorherigen Seite	
Symbol/Bezeichner	Bedeutung
$det(\boldsymbol{A})$	Die Determinante einer Matrix $\boldsymbol{A}$
$SVD(\boldsymbol{A})$	Die Singulärwertzerlegung einer Matrix $\boldsymbol{A}$: $SVD(\boldsymbol{A}) = \boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}}$
N	Die maximale Anzahl an Iterationen eines RANSAC-Durchlaufs
p	Die Wahrscheinlichkeit, dass RANSAC mindestens einen ausreißerfreien Datensatz zieht
w	Das Verhältnis der geschätzten Inlier der RANSAC-Messdatenmenge
n	Die minimale Anzahl an 2D-3D-Korrespondenzen, die für eine Modellschätzung nötig sind
τ	Die maximal erlaubte RANSAC-Distanz

Tabelle A.1: Übersicht über die verwendeten Symbole und Bezeichner.

Anhang B

Mathematische Verfahren

B.1 Direct Linear Transform

Gegeben: Punkt $\begin{bmatrix} x_n & y_n \end{bmatrix}^{\mathrm{T}}$ in Bildkoordinaten und Punkt $\begin{bmatrix} {X_n}^W & {Y_n}^W & {Z_n}^W \end{bmatrix}^{\mathrm{T}}$ in Weltkoordinaten.

Gesucht: Rotation $\boldsymbol{R}$ und Translation $\boldsymbol{t}$ der Transformation.

Der Bildpunkt muss auf einer Geraden zwischen Kameraursprung und Weltpunkt in Kamerakoordinaten liegen, dies ist das *Kollinearitätsprinzip* der DLT:

$$\lambda \begin{bmatrix} x_n \\ y_n \end{bmatrix} = \begin{bmatrix} {X_n}^C \\ {Y_n}^C \\ {Z_n}^C \end{bmatrix} \tag{B.1}$$

und somit ist das Kreuzprodukt zwischen Bildpunkt und Weltpunkt Null:

$$\begin{bmatrix} x_n \\ y_n \end{bmatrix} \times \begin{bmatrix} {X_n}^C \\ {Y_n}^C \\ {Z_n}^C \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{B.2}$$

Durch die Kreuzproduktmatrix ergibt sich:

$$\begin{bmatrix} 0 & -1 & y_n \\ 1 & 0 & -x_n \\ -y_n & x_n & 0 \end{bmatrix} \begin{bmatrix} R & t \end{bmatrix} \begin{bmatrix} {X_n}^W \\ {Y_n}^W \\ {Z_n}^W \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{B.3}$$

Es wird definiert:

$$\boldsymbol{P} \in \mathfrak{R}^{12x1} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 & r_{21} & r_{22} & r_{23} & t_2 & r_{31} & r_{32} & r_{33} \end{bmatrix}^{\mathrm{T}} \tag{B.4}$$

und

$$Q_n = \begin{bmatrix} {X_n}^W & {Y_n}^W & {Z_n}^W & 1 \end{bmatrix}^{\mathrm{T}} \tag{B.5}$$

Durch Umsortieren ergibt sich:

$$\begin{bmatrix} 0000 & -{Q_n}^{\mathrm{T}} & y_n{Q_n}^{\mathrm{T}} \\ {Q_n}^{\mathrm{T}} & 0000 & -x_n{Q_n}^{\mathrm{T}} \\ -y_n{Q_n}^{\mathrm{T}} & x_n{Q_n}^{\mathrm{T}} & 0000 \end{bmatrix}_{3\times 12} \boldsymbol{P} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{B.6}$$

Für jeden Punkt entstehen also 3 Gleichungen, wovon nur 2 linear unabhängig sind. Der Parameterraum enthält 12 Unbekannte und folglich werden zur Lösung des Gleichungssystems $n \geq 6$ Punktkorrespondenzen benötigt.

$$\boldsymbol{A} = \begin{bmatrix} {Q_1}^{\mathrm{T}} & 0000 & -x_1{Q_1}^{\mathrm{T}} \\ 0000 & {Q_1}^{\mathrm{T}} & -y_1{Q_1}^{\mathrm{T}} \\ \vdots & \vdots & \vdots \\ {Q_n}^{\mathrm{T}} & 0000 & -x_n{Q_n}^{\mathrm{T}} \\ 0000 & {Q_n}^{\mathrm{T}} & -y_n{Q_n}^{\mathrm{T}} \end{bmatrix} \tag{B.7}$$

Die so erstellte Messwertmatrix $\boldsymbol{A}$ hat den Rang 11 und $\boldsymbol{P}$ spannt ihren Nullraum auf:

$$\boldsymbol{A}\boldsymbol{P} = 0 \tag{B.8}$$

Die Lösung für $\boldsymbol{P}$ ist bis auf einen Skalierungsfaktor eindeutig mittels Singulärwertzerlegung zu bestimmen:

$$SVD(\boldsymbol{A}) = \boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}} \tag{B.9}$$

$\boldsymbol{A}$ hat den Rang 11 und $\boldsymbol{P}$ korrespondiert mit dem Vektor zum einzigen Singulärwert 0, die letzte Spalte von $\boldsymbol{V}$. Der Skalierungsfaktor s errechnet sich aus:

$$s = det(\begin{bmatrix} P_1 & P_2 & P_3 \\ P_5 & P_6 & P_7 \\ P_9 & P_{10} & P_{11} \end{bmatrix})^{\frac{1}{3}} \tag{B.10}$$

Im Falle einer negativen Determinante:

$$s = -1(-1det(\begin{bmatrix} P_1 & P_2 & P_3 \\ P_5 & P_6 & P_7 \\ P_9 & P_{10} & P_{11} \end{bmatrix}))^{\frac{1}{3}} \tag{B.11}$$

$\boldsymbol{R}$ und $\boldsymbol{t}$ ergeben sich aus den skalierten Werten von $\boldsymbol{P}$.

B.2 Bestimmung der nächsten Rotationsmatrix

Im Folgenden werden die Eigenschaften einer Rotationsmatrix $\boldsymbol{R}$ nach [TV98] beschrieben.

Für eine Rotationsmatrix $\boldsymbol{R}$ gilt:

$$\boldsymbol{R}\,\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{R}^{\mathrm{T}}\,\boldsymbol{R} = \boldsymbol{I} \tag{B.12}$$

wobei $\boldsymbol{I}$ die Einheitsmatrix bezeichnet. B.12 beschreibt die Orthogonalität einer Rotationsmatrix.

$$det(\boldsymbol{R}) = 1 \tag{B.13}$$

zeigt die Eigenschaft einer Rotationsmatrix, dass sie die relative Orientierung der Referenzkoordinatensysteme erhält, also deren rechtshändische oder linkshändische Ausrichtung.

$$\sum_{j=1}^{3} r_{ij} r_{ik} = \sum_{j=1}^{3} r_{ji} r_{jk} = \begin{cases} 0 & i \neq k \\ 1 & i = k \end{cases} \tag{B.14}$$

lässt sich aus der Ansicht, dass ein Eintrag r_{ij} einer Rotationsmatrix dem Kosinus des Winkels zwischen den Basisvektoren der jeweilig zu rotierenden Koordinatensysteme entspricht, veranschaulichen. Die Spalten und Zeilen von $\boldsymbol{R}$ sind demnach orthogonale Einheitsvektoren.

Zu einer beliebigen 3×3 Matrix $\boldsymbol{A}$ lässt sich die nächste[1] Rotationsmatrix $\boldsymbol{R}$ wie folgt bestimmen.

Zerlegung der Matrix $\boldsymbol{A}$ mittels Singulärwertzerlegung:

$$SVD(\boldsymbol{A}) = \boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}} \tag{B.15}$$

Die Singulärwerte σ einer Rotationsmatrix sind 1 und die Matrix hat vollen Rang. Die nächste Rotationsmatrix zu $\boldsymbol{A}$ ist also definiert durch:

$$\boldsymbol{R} = \boldsymbol{U} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{V}^{\mathrm{T}} \tag{B.16}$$

Somit ist $det(\boldsymbol{R}) = 1$, $\boldsymbol{R}$ ist orthonormal und erfüllt damit alle geforderten Eigenschaften einer Rotationsmatrix.

[1] Im Sinne der Frobenius-Norm

Anhang C

Implementation

Nachfolgend werden einige Implementationsdetails der Arbeit aufgezeigt.

C.1 Benutzeroberfläche

Die gesamte Systemfunktionalität wurde aus bedienungsfreundlichen Gründen in ein graphisches Benutzerinterface integriert.

Die Benutzeroberfläche bietet eine Registrierkarte zur Ansicht der Modellerstellung und eine Registrierkarte für die Poseschätzung, deren Funktionalität in C.1.1 und C.1.2 beschrieben wird.

C.1.1 Stereoansicht

Diese Registrierkarte bedient die Anforderungen an einen Stereoalgorithmus. Es kann eine Bildserie geladen werden, deren einzelne Bilder in einem Drop-Down-Menu selektiert werden können. Auf jedem Bildpaar wird danach der SIFT-Algorithmus aufgerufen, dessen Ergebnis in Form von roten Punkten (den SIFT-Features) auf das originale Bild überlagert wird. Aus den in 4.1.4 erwähnten Gründen ist es in dieser Arbeit erforderlich, manuell Korrespondenzen zwischen den fortlaufenden Bildern zu bestimmen. Die SIFT-Merkmale können also per Mausklick selektiert werden und erscheinen farblich paarweise unterschieden in der Bildansicht.

Die Parameter zur nächsten Nachbarschaftssuche und der epipolare Schwellwert können ebenfalls dynamisch angepasst werden. Darüber hinaus besteht die Möglichkeit, zwischen den verschiedenen Modellerstellungsverfahren zu wählen und das Modell daraufhin zu triangulieren.

Neben einer Zoom-Funktion und der Wahl einer Kamerakalibrierungsdatei bietet diese Ansicht die Option, sämtliche bisher getätigten Einstellungen und Zustände in Dateien zu speichern und auch wieder zu laden.

Abbildung C.1 zeigt die Modellansicht in Aktion. Die grün dargestellten Merkmale sind dabei konsistente Merkmale, die zur Erstellung des dreidimensionalen Modells verwendet werden. Aus visuellen Evaluierungsgründen ist jeweils deren korrespondierende Nummer annotiert.

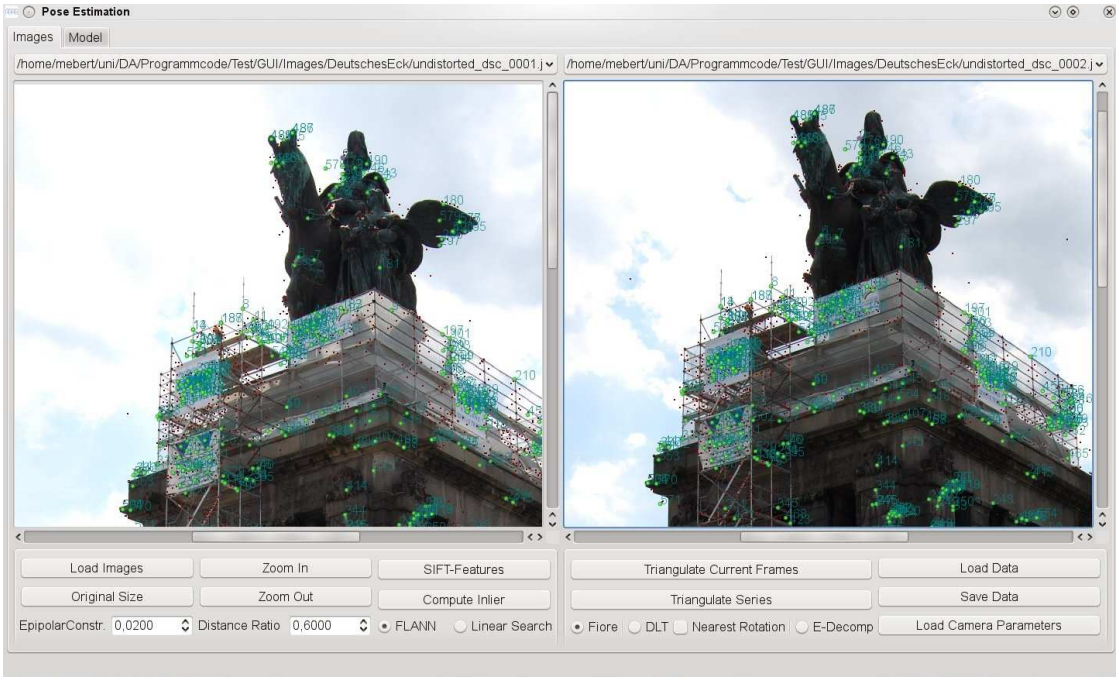


Abbildung C.1: Die Stereoansicht der Benutzeroberfläche

C.1.2 Modellansicht

Die als *Model* bezeichnete Registrierkarte ist zur Visualisierung der gewonnenen dreidimensionalen Information und der Schätzung einer neuen Kamerapose bestimmt. In einem mit *OpenGL* entwickelten Fenster werden die von der Modellerstellung gewonnenen Daten angezeigt. Sie beinhalten die Weltpunkte und die Kamerapositionen der Referenzbilder in Magenta dargestellt. Der Ursprung des Weltkoordinatensystems ist deckungsgleich mit der ersten Kameraposition und dessen Basisvektoren sind farblich hervorgehoben. Eine Steuerung der Kamerasicht und eine Funktion, sich mittels Tastendruck in die neu geschätzte Kamerapose zu begeben ist hier ebenfalls integriert.

Die rechte Seite bildet das Bedienfeld für die Poseschätzung. Hier kann ein Bild sowie die Kamerakalibrierdatei der Posekamera[1] geladen werden. Nach einem erneuten Aufruf des SIFT-Algorithmus werden ebenfalls in dem Posebild die gefundenen Merkmale rot eingezeichnet. Neben den Schwellwertreglern für den maximal erlaubten Rückprojektionsfehler, dem erwarteten Inlier-Verhältnis von RANSAC und der bekannten Distance-Ration findet sich ein Knopf zur Auslösung der Poseschätzung.

Die neu geschätzte Kamerapose erscheint als grün gezeichneter Kameraframe in der dreidimensionalen Ansicht (in Abbildung C.2 ist die Position der geschätzten Kamera vor der ersten Treppenreihe zu finden). Die zur Schätzung der Kamerapose verwendeten 2D-3D-Korrespondenzen werden sowohl im Modell als auch im Posebild mittels grüner Quadrate dargestellt. Die rückprojizierten Punkte aus der aktuell geschätzten Kamerapose und den verwendeten Weltpunkten werden als optisches Überprüfungsmittel nochmals in das Posebild als blaue halbtransparente Kreise markiert.

Das Ergebnis der Poseschätzung in Form der Kameramatrix wird in der 3×4 Matrixdarstellung im rechten unteren Bereich ausgegeben.

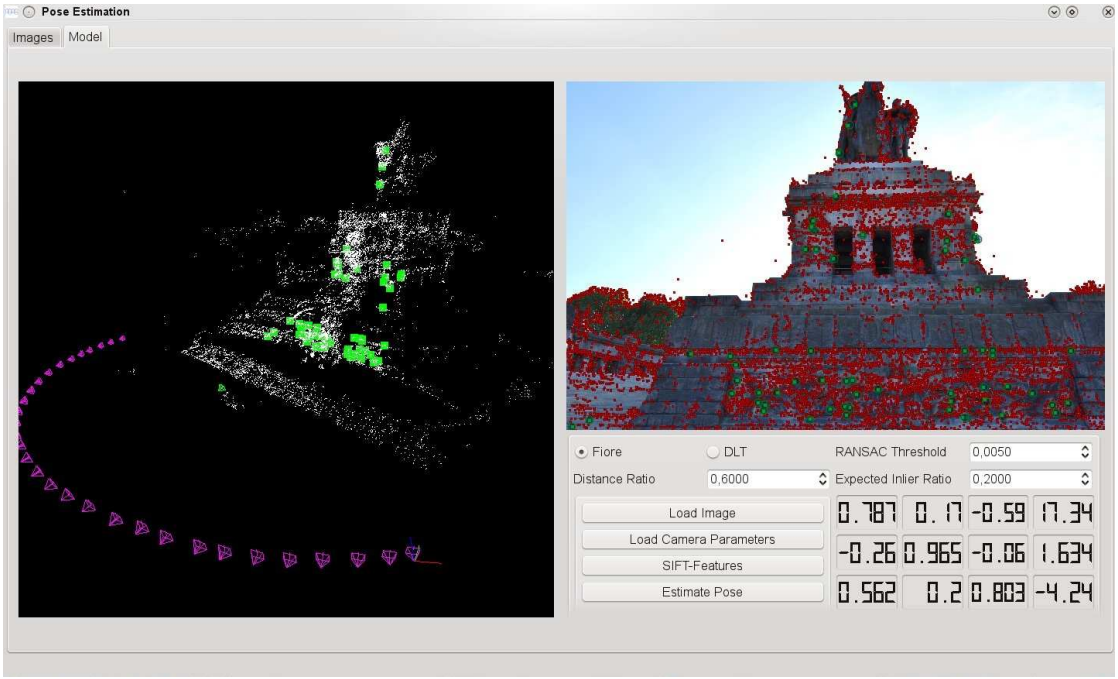


Abbildung C.2: Die Modellansicht der Benutzeroberfläche

[1] Falls abweichend von der Kamera, die zur Modellerstellung verwendet wurde.

C.2 Datenverwaltung

Während der Modellerstellung wird eine große Datenmenge generiert, die es zu verwalten gilt. Sie setzt sich zusammen aus den extrahierten Bildmerkmalen, den Weltpunkten des Modells, den zugehörigen Kameras der Bildserie und sämtliche in dieser Arbeit aufgezeigten Beziehungen unter ihnen. Diese Daten sind an den verschiedenen Positionen der Modellerstellung und während der Poseschätzung stets verfügbar zu halten.

In Hinsicht auf die Wiederverwendbarkeit des entstandenen Programmcodes ist eine zentrale Klasse `DataManager` implementiert, die sämtliche Daten als Membervariablen hält. Sie bietet den restlichen Klassen die nötigen Schnittstellen für Datenzugriffe an. Manipulierte oder neu erstellte Daten werden ebenfalls über die passenden Schnittstellen dem `DataManager` zurückgegeben.

Um Seiteneffekte durch Mehrfachzugriffe auf die Variablen zu vermeiden, ist der `DataManager` als *Singleton* umgesetzt. Zur Laufzeit des Programms existiert also stets nur eine Instanz des `DataManager`.

Darüber hinaus bietet diese Klasse die Möglichkeit, sämtliche Members in Binärdateien zu serialisieren und somit den Zustand der Daten zu speichern. Nach einem Deserialisierungsschritt liegen die geladenen Daten aus den Binärdateien wieder vor, ohne die Modelle komplett neu zu berechnen. Dies führt neben der Zeitersparnis zu einer guten Wiederverwendbarkeit der entstandenen Modelle.

Die 2D-3D-Korrespondenzen der Bildpunkte und der Weltpunkte sind in einer *Map*-Struktur gehalten. Die *Map* bietet die Möglichkeit der Mehrfachzuordnung von *Keys* und *Values*[2]. Dieses Konzept entspricht der Zuordnung mehrerer Bildpunkte aus verschiedenen Ansichten zu einem physischen Weltpunkt. Die Merkmalspunkte sind dabei die *Keys* und die Weltpunkte die *Values* der *Map*. Eine solche Datenstruktur ist für die Suche der passenden Werte eines Schlüssels optimiert, was der Suche eines Weltpunktes nach einem Bildpunkt gleichbedeutend ist. Die Mehrzahl der Suchanfragen wird sich in dieser Reihenfolge gestalten. Es ist aber ebenfalls möglich eine entgegengesetzte Suchanfrage zu stellen, wenn zum Beispiel von Interesse ist, wieviele Bildpunkte einen Weltpunkt beschreiben.

Dies ist in Hinblick auf weitere Optimierungen nützlich.

C.3 Benutzte Bibliotheken

Die Implementierung ist in *C++* unter GNU/Linux umgesetzt, die graphische Benutzeroberfläche ist mit *Qt* in der Version 4.5.3 entwickelt. Als build-Werkzeug wird das von *Qt* mitgelieferte Programm *qmake* in der Version 2.0.1a verwendet.

[2]Die Schlüssel der Map sind dabei wohl unterscheidbar.

Für lineare Algebra Routinen werden die Bibliotheken *lapackpp* v2.5.2 und *TooN* v2.0.0-beta4 benutzt.

Bei der Visualisierung der Modelle kommt *OpenGL* zum Einsatz. Für die Kalibrierung der Kamera sowie für die radiale Entzerrung der Bilder wird *OpenCV* in der aktuellen Version 2.0.0 benötigt.

Anhang D

Inhalt der DVD

```
Ausarbeitung/
  Arbeit/
  VortragOberseminar/
  Ebert2009.pdf

Literatur/

Programmcode/

Sonstiges/
  Modelldaten/
    DeutschesEck/
    Sparkassenhaus/
    Pforte/
    KaiserinAugusta/
    Campusmodell/
  Bilddaten/
    DeutschesEck/
    Sparkassenhaus/
    Pforte/
    KaiserinAugusta/
    Campusmodell/
```

Die Verzeichnisstruktur der beigelegten DVD ist wie folgt aufgebaut:

Im Ordner `Ausarbeitung` finden sich neben diesem Dokument (Ebert2009.pdf) dessen LaTeX-Quellen sowie die Quelldateien des Oberseminar-Vortrags vom 12. November 2009. Der Ordner `Literatur` enthält sämtliche in dieser Arbeit als Literatur angegebenen Dokumente, außer sie waren nur in Buchform vorliegend.

Die Literatureinträge können ebenfalls über die Literaturdatenbank der AGAS abgefragt werden.

Der Programmcode liegt im entsprechenden Verzeichnis als *C++*-Quellcode vor. Zu dessen Kompilierung ist lediglich ein Aufruf von *qmake* gefolgt von *make* und eine eventuelle Anpassung der Projektdatei nötig.

Im Ordner `Sonstiges` liegen sowohl Modelldaten als auch Bilddaten bereit. Bilddaten dienen zur Generierung eigener Modelle sowie zur Schätzung der Kamerapose. Die Modelldaten entsprechen den aus der Anwendung serialisierten Dateien und können somit in der Benutzeroberfläche geladen werden. Eine manuelle Erstellung der Modelle ist überflüssig, es kann direkt in die Modellansicht gewechselt werden und die Poseschätzung erfolgen. Außerdem finden sich in diesem Verzeichnis die von der Anwendung benötigten Kalibrierdateien.

Literaturverzeichnis

[BYY+08] Bastanlar, Y. ; Yilmaz, E. ; Yardimci, Y. ; Grammalidis, N. ; Zabulis, X. ; Triantafyllidis, G.: 3D Reconstruction for a Cultural Heritage Virtual Tour System, 2008, S. B5: 1023 ff

[Fau93] Faugeras, Olivier D.: *Three-Dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, Massachusetts : MIT Press, 1993

[FB81] Fischler, Martin A. ; Bolles, Robert C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. In: *Commun. ACM* 24 (1981), Nr. 6, S. 381–395

[Fio01] Fiore, Paul D.: Efficient linear solution of exterior orientation. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23 (2001), Nr. 2, S. 140–148

[FTZ99] Fitzgibbon, Andrew W. ; Torr, Philip H. ; Zisserman, Andrew: *The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences*. 1999

[GIM99] Gionis, Aristides ; Indyk, Piotr ; Motwani, Rajeev: Similarity Search in High Dimensions via Hashing. In: *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1999, S. 518–529

[GL06] Gordon, Iryna ; Lowe, David G.: What and Where: 3D Object Recognition with Accurate Pose. In: *Toward Category-Level Object Recognition*, Springer Verlag, 2006, S. 67–82

[Har98] Hartley, Richard I.: Minimizing Algebraic Error in Geometric Estimation Problems. In: *ICCV*, 1998, S. 469–476

[HS88] Harris, C. ; Stephens, M.: A combined corner and edge detector. In: *Fourth Alvey Vision Conference*. Manchester, UK, 1988, S. 147–151

[HZ03] Hartley, Richard I. ; Zisserman, Andrew: *Multiple View Geometry in Computer Vision*. 2. Cambridge University Press, 2003

[KL81] Kanade, Takeo ; Lucas, Bruce: An Iterative Image Registration Technique with an Application to Stereo Vision. In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, 1981, S. 674–679

[LF05] Lepetit, V. ; Fua, P.: Monocular Model-Based 3D Tracking of Rigid Objects: A Survey. In: *Foundations and Trends in Computer Graphics and Vision* 1 (2005), Nr. 1, S. 1–89

[Lin94] Lindeberg, Tony: *Scale-Space Theory in Computer Vision*. Dordrecht, Netherlands : Kluwer Academic Publishers, 1994

[Low99] Lowe, David G.: Object Recognition from Local Scale-Invariant Features. In: *Proceedings of the International Conference on Computer Vision*. Corfu Greece, 1999, S. 1150–1157

[Low04] Lowe, David G.: Distinctive Image Features from Scale-Invariant Keypoints. In: *International Journal of Computer Vision* 60 (2004), Nr. 2, S. 91–110

[ML09] Muja, Marius ; Lowe, David G.: Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In: *VISAPP 2009 - Proceedings of the Fourth International Conference on Computer Vision Theory and Applications, Lisboa, Portugal, February 5-8, 2009 - Volume 1*, INSTICC Press, 2009, S. 331–340

[MNT04] Madsen, K. ; Nielsen, H. B. ; Tingleff, O.: *Methods for Non-Linear Least Squares Problems*. 2004

[QL99] Quan, Long ; Lan, Zhongdan: Linear N-Point Camera Pose Determination. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21 (1999), Nr. 8, S. 774–780

[SZ02] Schaffalitzky, Frederik ; Zisserman, Andrew: Multi-view Matching for Unordered Image Sets, or "How Do I Organize My Holiday Snaps¿'. In: *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I*. London, UK : Springer-Verlag, 2002, S. 414–431

[TMHF00] Triggs, Bill ; McLauchlan, Philip ; Hartley, Richard I. ; Fitzgibbon, Andrew: Bundle Adjustment – A Modern Synthesis. In: Triggs, Bill (Hrsg.) ; Zisserman, Andrew (Hrsg.) ; Szeliski, Richard (Hrsg.): *Vision Algorithms: Theory and Practice*. Springer Verlag, 2000, S. 298–375

[TV98] Trucco, E. ; Verri, A.: *Introductory Techniques for 3-D Computer Vision*. New York : Prentice Hall, 1998

[VL04] Vacchetti, Luca ; Lepetit, Vincent: Stable Real-Time 3D Tracking Using Online and Offline Information. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26 (2004), Nr. 10, S. 1385–1391. – Member-Pascal Fua

[Zha00] Zhang, Zhengyou: A flexible new technique for camera calibration. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (2000), Nr. 11, S. 1330–1334